CERN 99–03
19 July 1999

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

BAYESIAN REASONING IN HIGH-ENERGY PHYSICS: PRINCIPLES AND APPLICATIONS

G. D'Agostini

GENEVA
1999

ISSN 0007–8328
ISBN 92–9083–145–6

CERN 99–03
19 July 1999

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

BAYESIAN REASONING IN HIGH-ENERGY PHYSICS: PRINCIPLES AND APPLICATIONS

G. D'Agostini
Dip. Di Fisica, Università "La Sapienza", Roma, Italy and CERN, Geneva, Switzerland

GENEVA
1999

CERN–Service d'information scientifique–RD/990–2500–juillet 1999

Abstract

Bayesian statistics is based on the intuitive idea that probability quantifies the degree of belief in the occurrence of an event. The choice of name is due to the key role played by Bayes' theorem, as a logical tool to update probability in the light of new pieces of information. This approach is very close to the intuitive reasoning of experienced physicists, and it allows all kinds of uncertainties to be handled in a consistent way. Many cases of evaluation of measurement uncertainty are considered in detail in this report, including uncertainty arising from systematic errors, upper/lower limits and unfolding. Approximate methods, very useful in routine applications, are provided and several standard methods are recovered for cases in which the (often hidden) assumptions on which they are based hold.

Contents

Introduction

These notes are based on seminars and minicourses given in various places over the last four years. In particular, lectures I gave to graduate students in Rome and to summer students in DESY in the spring and summer of 1995 encouraged me to write the 'Bayesian primer', which still forms the core of this script. I took advantage of the academic training given at CERN at the end of May 1998 to add some material developed in the meantime.

Instead of completely rewriting the primer, producing a thicker report which would have been harder to read sequentially, I have divided the text into three parts.

- The first part is dedicated to a critical review of standard statistical methods and to a general overview of the proposed alternative. It contains references to the other two parts for details.
- The second part essentially reproduces the old primer, subdivided into chapters for easier reading and with some small corrections.
- Part three contains an appendix, covering remarks on the general aspects of probability, as well as other applications.

The advantage of this structure is that the reader can have an overall view of problems and proposed solutions and then decide if he wants to enter into details.

This structure inevitably leads to some repetition, which I have tried to keep to a minimum. In any case, *repetita juvant*, especially in this subject where the real difficulty is not understanding the formalism, but shaking off deep-rooted prejudices. This is also the reason why this report is somewhat verbose (I have to admit) and contains a plethora of footnotes, indicating that this topic requires a more extensive treatise.

A last comment concerns the title of the report. As discussed in the last lecture at CERN, a title which was closer to the spirit of the lectures would have been "Probabilistic reasoning ... ". In fact, I think the important thing is to have a theory of uncertainty in which "probability" has the same meaning for everybody: precisely that meaning which the human mind has developed naturally and which frequentists have tried to kill. Using the term "Bayesian" might seem somewhat reductive, as if the methods illustrated here would always require explicit use of Bayes' theorem. However, in common usage 'Bayesian' is a synonym of 'based on subjective probability', and this is the reason why these methods are the most general to handle uncertainty. Therefore, I have left the title of the lectures, with the hope of attracting the attention of those who are curious about what 'Bayesian' might mean.

Email: dagostini@roma1.infn.it
URL: http://www-zeus.roma1.infn.it/~agostini/

Part I

Subjective probability in physics? Scientific reasoning in conditions of uncertainty

Chapter 1

Uncertainty in physics and the usual methods of handling it

"In almost all circumstances, and at all times,
we find ourselves in a state of uncertainty.
Uncertainty in every sense.
Uncertainty about actual situations, past and present ...
Uncertainty in foresight: this would not be eliminated
or diminished even if we accepted, in its most absolute form,
the principle of determinism; in any case, this is no longer in fashion.
Uncertainty in the face of decisions: more than ever in this case ...
Even in the field of tautology (i.e of what is true or false by mere
definition, independently of any contingent circumstances) we always
find ourselves in a state of uncertainty ... (for instance,
of what is the seventh, or billionth, decimal place of π *...) ... "*
(Bruno de Finetti)

1.1 Uncertainty in physics

It is fairly well accepted among physicists that any conclusion which results from a measurement is affected by a certain degree of uncertainty. Let us remember briefly the reasons which prevent us from reaching certain statements. Figure 1.1 sketches the activity of physicists (or of any other

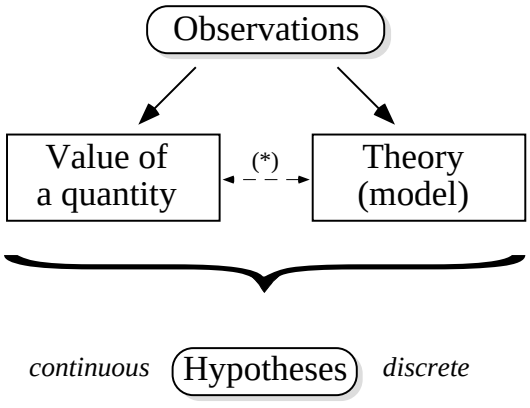

Figure 1.1: From observations to hypotheses. The link between value of a quantity and theory is a reminder that sometimes a physics quantity has meaning only within a given theory or model.

scientist). From experimental data one wishes to determine the value of a given quantity, or to establish which theory describes the observed phenomena better. Although they are often seen as separate, both tasks may be viewed as two sides of the same process: going from observations to hypotheses. In fact, they can be stated in the following terms.

A: Which values are (more) compatible with the definition of the measurand, under the condition that certain numbers have been observed on instruments (and subordinated to all the available knowledge about the instrument and the measurand)?

B: Which theory is (more) compatible with the observed phenomena (and subordinated to the credibility of the theory, based also on aesthetics and simplicity arguments)?

The only difference between the two processes is that in the first the number of hypotheses is virtually infinite (the quantities are usually supposed to assume continuous values), while in the second it is discrete and usually small.

The reasons why it is impossible to reach the ideal condition of certain knowledge, i.e. only one of the many hypotheses is considered to be true and all the others false, may be summarized in the following, well-understood, scheme.

A: As far as the determination of the value of a quantity is concerned, one says that *"uncertainty is due to measurement errors"*.

B: In the case of a theory, we can distinguish two subcases:

(B_1) The law is probabilistic, i.e. the observations are not just a logical consequence of the theory. For example, tossing a regular coin, the two sequences of heads and tails

`hhhhhhhhhhhhhhhhhhhhhhhhh`

`hhttttthhhhthhtthhhthtthht`

have the same probability of being observed (as any other sequence). Hence, there is no way of reaching a firm conclusion about the regularity of the coin after an observed sequence of any particular length.[1]

(B_2) The law is deterministic. But this property is only valid in principle, as can easily be understood. In fact, in all cases the actual observations also depend on many other factors external to the theory, such as initial and boundary conditions, influence factors, experimental errors, etc. All unavoidable uncertainties on these factors mean that the link between theory and observables is of a probabilistic nature in this case too.

1.2 True value, error and uncertainty

Let us start with case **A**. A first objection would be *"What does it mean that uncertainties are due to errors? Isn't this just tautology?"*. Well, the nouns 'error' and 'uncertainty', although currently used almost as synonyms, are related to different concepts. This is a first hint that in this subject there is neither uniformity of language, nor of methods. For this reason the metrological organizations have recently made great efforts to bring some order into the field [1, 2, 3, 4, 5].

[1] But after observation of the first sequence one would strongly suspect that the coin had two heads, if one had no means of directly checking the coin. The concept of probability will be used, in fact, to quantify the degree of such suspicion.

In particular, the International Organization for Standardization (ISO) has published a *"Guide to the expression of uncertainty in measurement"* [3], containing definitions, recommendations and practical examples. Consulting the 'ISO Guide' we find the following definitions.

- Uncertainty: *"a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurement."*
- Error: *"the result of a measurement minus a true value of the measurand."*

One has to note the following.

- The ISO definition of uncertainty defines the concept; as far as the operative definition is concerned, they recommend the 'standard uncertainty', i.e. the standard deviation (σ) of the possible values that the measurand may assume (each value is weighted with its 'degree of belief' in a way that will become clear later).
- It is clear that the error is usually unknown, as follows from the definition.
- The use of the article 'a' (instead of 'the') when referring to 'true value' is intentional, and rather subtle.

Also the ISO definition of true value differs from that of standard textbooks. One finds, in fact:

- true value: *"a value compatible with the definition of a given particular quantity."*

This definition may seem vague, but it is more practical and pragmatic, and of more general use, than *"the value obtained after an infinite series of measurements performed under the same conditions with an instrument not affected by systematic errors."* For instance, it holds also for quantities for which it is not easy to repeat the measurements, and even for those cases in which it makes no sense to speak about repeated measurements under the same conditions. The use of the indefinite article in conjunction with true value can be understood by considering the first item on the list in the next section.

1.3 Sources of measurement uncertainty

It is worth reporting the sources of uncertainty in measurement as listed by the ISO Guide:

"1 incomplete definition of the measurand;

2 imperfect realization of the definition of the measurand;

3 non-representative sampling – the sample measured may not represent the defined measurand;

4 inadequate knowledge of the effects of environmental conditions on the measurement, or imperfect measurement of environmental conditions;

5 personal bias in reading analogue instruments;

6 finite instrument resolution or discrimination threshold;

7 inexact values of measurement standards and reference materials;

8 inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;

9 approximations and assumptions incorporated in the measurement method and procedure;

10 variations in repeated observations of the measurand under apparently identical conditions."

These do not need to be commented upon. Let us just give examples of the first two sources.

1. If one has to measure the gravitational acceleration g at sea level, without specifying the precise location on the earth's surface, there will be a source of uncertainty because many different — even though 'intrinsically very precise' — results are consistent with the definition of the measurand.[2]

2. The magnetic moment of a neutron is, in contrast, an unambiguous definition, but there is the experimental problem of performing experiments on isolated neutrons.

In terms of the usual jargon, one may say that sources 1–9 are related to systematic effects and 10 to 'statistical effects'. Some caution is necessary regarding the sharp separation of the sources, which is clearly somehow artificial. In particular, all sources 1–9 may contribute to 10, because they each depend upon the precise meaning of the clause *"under apparently identical conditions"* (one should talk, more precisely, about 'repeatability conditions' [3]). In other words, if the various effects change during the time of measurement, without any possibility of monitoring them, they contribute to the random error.

1.4 Usual handling of measurement uncertainties

The present situation concerning the treatment of measurement uncertainties can be summarized as follows.

- Uncertainties due to statistical errors are currently treated using the frequentistic concept of 'confidence interval', although
  - there are well known cases — of great relevance in frontier physics — in which the approach is not applicable (e.g. small number of observed events, or measurement close to the edge of the physical region);
  - the procedure is rather unnatural, and in fact the interpretation of the results is unconsciously subjective (as will be discussed later).
- There is no satisfactory theory or model to treat uncertainties due to systematic errors[3] consistently. Only *ad hoc* prescriptions can be found in the literature and in practice (*"my supervisor says ... "*): *"add them linearly"; "add them linearly if ... , else add them quadratically"; "don't add them at all".*[4] The fashion at the moment is to add them quadratically if they are considered to be independent, or to build a covariance matrix of

[2] It is then clear that the definition of true value implying an indefinite series of measurements with ideal instrumentation gives the illusion that the true value is unique. The ISO definition, instead, takes into account the fact that measurements are performed under real conditions and can be accompanied by all the sources of uncertainty in the above list.

[3] To be more precise one should specify 'of unknown size', since an accurately assessed systematic error does not yield uncertainty, but only a correction to the raw result.

[4] By the way, it is a good and recommended practice to provide the complete list of contributions to the overall uncertainty [3]; but it is also clear that, at some stage, the producer or the user of the result has to combine the uncertainty to form his idea about the interval in which the quantity of interest is believed to lie.

statistical and systematic contribution to treat the general case. In my opinion, besides all the theoretically motivated excuses for justifying this praxis, there is simply the reluctance of experimentalists to combine linearly 10, 20 or more contributions to a global uncertainty, as the (out of fashion) 'theory' of maximum bounds would require.[5]

The problem of interpretation will be treated in the next section. For the moment, let us see why the use of standard propagation of uncertainty, namely

$$\sigma^2(Y) = \sum_i \left(\frac{\partial Y}{\partial X_i}\right)^2 \sigma^2(X_i) + \textit{correlation terms}\,, \tag{1.1}$$

is not justified (especially if contributions due to systematic effects are included). This formula is derived from the rules of probability distributions, making use of linearization (a usually reasonable approximation for routine applications). This leads to theoretical and practical problems.

- X_i and Y should have the meaning of random variables.
- In the case of systematic effects, how do we evaluate the input quantities $\sigma(X_i)$ entering in the formula in a way which is consistent with their meaning as standard deviations?
- How do we properly take into account correlations (assuming we have solved the previous questions)?

It is very interesting to go to your favourite textbook and see how 'error propagation' is introduced. You will realize that some formulae are developed for random quantities, making use of linear approximations, and then suddenly they are used for physics quantities without any justification.[6] A typical example is measuring a velocity $v \pm \sigma(v)$ from a distance $s \pm \sigma(s)$ and a time interval $t \pm \sigma(t)$. It is really a challenge to go from the uncertainty on s and t to that of v without considering s, t and v as random variables, and to avoid thinking of the final result as a probabilistic statement on the velocity. Also in this case, an intuitive interpretation conflicts with standard probability theory.

1.5 Probability of observables versus probability of true values

The criticism about the inconsistent interpretation of results may look like a philosophical quibble, but it is, in my opinion, a crucial point which needs to be clarified. Let us consider the example of n independent measurements of the same quantity under identical conditions (with n large enough to simplify the problem, and neglecting systematic effects). We can evaluate the arithmetic average $\overline{x}$ and the standard deviation σ. The result on the true value μ is

$$\mu = \overline{x} \pm \frac{\sigma}{\sqrt{n}}\,. \tag{1.2}$$

[5] And in fact, one can see that when there are only two or three contributions to the 'systematic error', there are still people who prefer to add them linearly.

[6] Some others, including some old lecture notes of mine, try to convince the reader that the propagation is applied to the observables, in a very complicated and artificial way. Then, later, as in the 'game of the three cards' proposed by professional cheaters in the street, one uses the same formulae for physics quantities, hoping that the students do not notice the logical gap.

The reader will have no difficulty in admitting that the large majority of people interpret (1.2) as if it were[7]

$$P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\% \,. \tag{1.3}$$

However, conventional statistics says only that[8]

$$P(\mu - \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\% \,, \tag{1.4}$$

a probabilistic statement about $\overline{X}$, given μ, σ and n. Probabilistic statements concerning μ are not foreseen by the theory (*"μ is a constant of unknown value"*[9]), although this is what we are, intuitively, looking for: Having observed the *effect* $\overline{x}$ we are interested in stating something about the possible true value responsible for it. In fact, when we do an experiment, we want to increase our knowledge about μ and, consciously or not, we want to know which values are more or less probable. A statement concerning the probability that an observed value falls within a certain interval around μ is meaningless if it cannot be turned into an expression which states the quality of the knowledge about μ itself. Since the usual probability theory does not help, the probability inversion is performed intuitively. In routine cases it usually works, but there are cases in which it fails (see Section 1.7).

1.6 Probability of the causes

Generally speaking, what is missing in the usual theory of probability is the crucial concept of probability of hypotheses and, in particular, probability of causes: *"the essential problem of the experimental method"* (Poincaré):

> *"I play at écarté with a gentleman whom I know to be perfectly honest. What is the chance that he turns up the king? It is 1/8. This is a problem of the probability of effects. I play with a gentleman whom I do not know. He has dealt ten times, and he has turned the king up six times. What is the chance that he is a sharper? This is a problem in the probability of causes. It may be said that it is the essential problem of the experimental method"* [6].
> *"... the laws are known to us by the observed effects. Trying to deduct from the effects the laws which are the causes, it is solving a problem of probability of causes"* [7].

A theory of probability which does not consider probabilities of hypothesis is unnatural and prevents transparent and consistent statements about the causes which may have produced the observed effects from being assessed.

[7] There are also those who express the result, making the trivial mistake of saying *"this means that, if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval* $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$*"*. (Besides the interpretation problem, there is a missing factor of $\sqrt{2}$ in the width of the interval ...)

[8] The capital letter to indicate the average appearing in (1.4) is used because here this symbol stands for a random variable, while in (1.3) it indicated a realization of it. For the Greek symbols this distinction is not made, but the different role should be evident from the context.

[9] It is worth noting the paradoxical inversion of role between μ, about which we are in a state of uncertainty, considered to be a constant, and the observation $\overline{x}$, which has a certain value and which is instead considered a random quantity. This distorted way of thinking produces the statements to which we are used, such as speaking of *"uncertainty (or error) on the observed number"*: If one observes 10 on a scaler, there is no uncertainty on this number, but on the quantity which we try to infer from the observation (e.g. λ of a Poisson distribution, or a rate).

1.7 Unsuitability of confidence intervals

According to the standard theory of probability, statement (1.3) is nonsense, and, in fact, good frequentistic books do not include it. They speak instead about 'confidence intervals', which have a completely different interpretation [that of (1.4)], although several books and many teachers suggest an interpretation of these intervals as if they were probabilistic statements on the true values, like (1.3). But it seems to me that it is practically impossible, even for those who are fully aware of the frequentistic theory, to avoid misleading conclusions. This opinion is well stated by Howson and Urbach in a paper to Nature [8]:

> *"The statement that such-and-such is a 95% confidence interval for μ seems objective. But what does it say? It may be imagined that a 95% confidence interval corresponds to a 0.95 probability that the unknown parameter lies in the confidence range. But in the classical approach, μ is not a random variable, and so has no probability. Nevertheless, statisticians regularly say that one can be '95% confident' that the parameter lies in the confidence interval. They never say why."*

The origin of the problem goes directly to the underlying concept of probability. The frequentistic concept of confidence interval is, in fact, a kind of artificial invention to characterize the uncertainty consistently with the frequency-based definition of probability. But, unfortunately – as a matter of fact – this attempt to classify the state of uncertainty (on the true value) trying to avoid the concept of probability of hypotheses produces misinterpretation. People tend to turn arbitrarily (1.4) into (1.3) with an intuitive reasoning that I like to paraphrase as 'the dog and the hunter': We know that a dog has a 50% probability of being 100 m from the hunter; if we observe the dog, what can we say about the hunter? The terms of the analogy are clear:

$$\begin{aligned} \text{hunter} &\leftrightarrow \text{true value} \\ \text{dog} &\leftrightarrow \text{observable}\,. \end{aligned}$$

The intuitive and reasonable answer is *"The hunter is, with 50% probability, within 100 m of the position of the dog."* But it is easy to understand that this conclusion is based on the tacit assumption that 1) the hunter can be anywhere around the dog; 2) the dog has no preferred direction of arrival at the point where we observe him. Any deviation from this simple scheme invalidates the picture on which the inversion of probability (1.4) $\rightarrow$ (1.3) is based. Let us look at some examples.

Example 1: Measurement at the edge of a physical region.

An experiment, planned to measure the electron-neutrino mass with a resolution of $\sigma = 2\,\mathrm{eV}/c^2$ (independent of the mass, for simplicity, see Fig. 1.2), finds a value of $-4\,\mathrm{eV}/c^2$ (i.e. this value comes out of the analysis of real data treated in exactly the same way as that of simulated data, for which a $2\,\mathrm{eV}/c^2$ resolution was found).

What can we say about m_ν?

$$m_\nu = -4 \pm 2\ \mathrm{eV}/c^2\ \ ?$$

$$P(-6\,\mathrm{eV}/c^2 \leq m_\nu \leq -2\ \mathrm{eV}/c^2) = 68\%\ \ ?$$

$$P(m_\nu \leq 0\ \mathrm{eV}/c^2) = 98\%\ \ ?$$

No physicist would sign a statement which sounded like he was 98% sure of having found a negative mass!

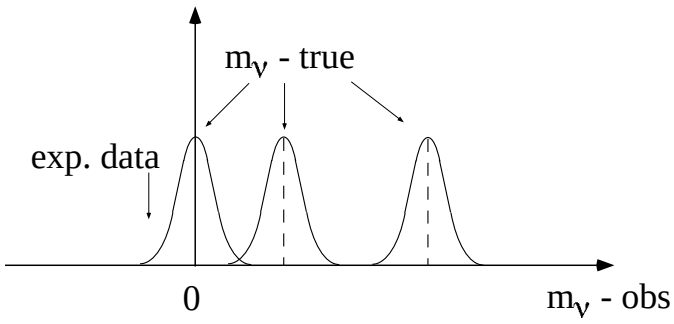

Figure 1.2: Negative neutrino mass?

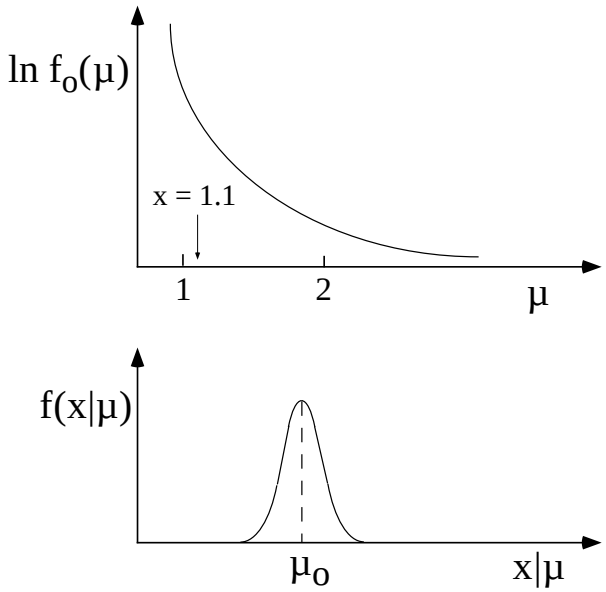

Figure 1.3: Case of highly asymmetric expectation on the physics quantity.

Example 2: Non-flat distribution of a physical quantity.

Let us take a quantity μ that we know, from previous knowledge, to be distributed as in Fig. 1.3. It may be, for example, the energy of bremsstrahlung photons or of cosmic rays. We know that an observable value X will be normally distributed around the true value μ, independently of the value of μ. We have performed a measurement and obtained $x = 1.1$, in arbitrary units. What can we say about the true value μ that has caused this observation? Also in this case the formal definition of the confidence interval does not work. Intuitively, we feel that there is more chance that μ is on the left side of (1.1) than on the right side. In the jargon of the experimentalists, *"there are more migrations from left to right than from right to left".*

Example 3: High-momentum track in a magnetic spectrometer.

The previous examples deviate from the simple dog–hunter picture only because of an asymmetric possible position of the 'hunter'. The case of a very-high-momentum track in a central detector of a high-energy physics (HEP) experiment involves asymmetric response of a detector for almost straight tracks and non-uniform momentum distribution of charged particles produced in the collisions. Also in this case the simple inversion scheme does not

work.

To sum up the last two sections, we can say that intuitive inversion of probability

$$P(\ldots \leq \overline{X} \leq \ldots) \Longrightarrow P(\ldots \leq \mu \leq \ldots)\,, \tag{1.5}$$

besides being theoretically unjustifiable, yields results which are numerically correct only in the case of symmetric problems.

1.8 Misunderstandings caused by the standard paradigm of hypothesis tests

Similar problems of interpretation appear in the usual methods used to test hypotheses. I will briefly outline the standard procedure and then give some examples to show the kind of paradoxical conclusions that one can reach.

A frequentistic hypothesis test follows the scheme outlined below (see Fig. 1.4). [10]

1. Formulate a hypothesis $H_\circ$.
2. Choose a test variable θ of which the probability density function $f(\theta \,|\, H_\circ)$ is known (analytically or numerically) for a given $H_\circ$.
3. Choose an interval $[\theta_1, \theta_2]$ such that there is high probability that θ falls inside the interval:

   $$P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha\,, \tag{1.6}$$

   with α typically equal to 1% or 5%.
4. Perform an experiment, obtaining $\theta = \theta_m$.
5. Draw the following conclusions :
   - if $\theta_1 \leq \theta_m \leq \theta_2 \quad \Longrightarrow H_\circ$ accepted;
   - otherwise $\quad \Longrightarrow H_\circ$ rejected with a significance level α.

The usual justification for the procedure is that the probability α is so low that it is practically impossible for the test variable to fall outside the interval. Then, if this event happens, we have good reason to reject the hypothesis.

One can recognize behind this reasoning a revised version of the classical 'proof by contradiction' (see, e.g., Ref. [10]). In standard dialectics, one assumes a hypothesis to be true and looks for a logical consequence which is manifestly false in order to reject the hypothesis. The slight difference is that in the hypothesis test scheme, the false consequence is replaced by an improbable one. The argument may look convincing, but it has no grounds. In order to analyse the problem well, we need to review the logic of uncertainty. For the moment a few examples are enough to indicate that there is something troublesome behind the procedure.

[10] At present, 'P-values' (or 'significance probabilities') are also *"used in place of hypothesis tests as a means of giving more information about the relationship between the data and the hypothesis than does a simple reject/do not reject decision"* [9]. They consist in giving the probability of the 'tail(s)', as also usually done in HEP, although the name 'P-values' has not yet entered our lexicon. Anyhow, they produce the same interpretation problems of the hypothesis test paradigm (see also example 8 of next section).

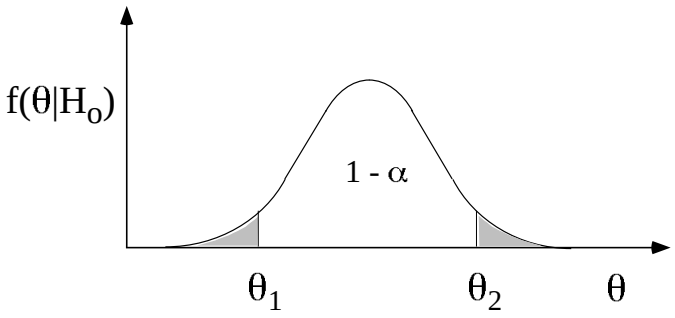

Figure 1.4: Hypothesis test scheme in the frequentistic approach.

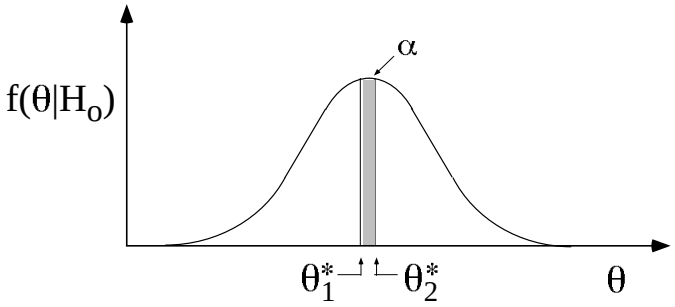

Figure 1.5: Would you accept this scheme to test hypotheses?

Example 4: Choosing the rejection region in the middle of the distribution.

Imagine choosing an interval $[\theta_1^*, \theta_2^*]$ around the expected value of θ (or around the mode) such that

$$P(\theta_1^* \leq \theta \leq \theta_2^*) = \alpha\,, \tag{1.7}$$

with α small (see Fig. 1.5). We can then reverse the test, and reject the hypothesis if the measured θ_m is inside the interval. This strategy is clearly unacceptable, indicating that the rejection decision cannot be based on the argument of practically impossible observations (smallness of α).

One may object that the reason is not only the small probability of the rejection region, but also its distance from the expected value. Figure 1.6 is an example against this objection. Although the situation is not as extreme as that depicted in Fig. 1.5, one would need a certain amount of courage to say that the $H_\circ$ is rejected if the test variable falls by chance in 'the bad region'.

Example 5: Has the student made a mistake?

A teacher gives to each student an individual sample of 300 random numbers, uniformly distributed between 0 and 1. The students are asked to calculate the arithmetic average. The prevision[11] of the teacher can be quantified with

$$\mathrm{E}\left[\overline{X}_{300}\right] = \frac{1}{2} \tag{1.8}$$

$$\sigma\left[\overline{X}_{300}\right] = \frac{1}{\sqrt{12}} \cdot \frac{1}{\sqrt{300}} = 0.017\,, \tag{1.9}$$

[11] By prevision I mean, following [11], a probabilistic 'prediction', which corresponds to what is usually known as expectation value (see Section 5.2.2).

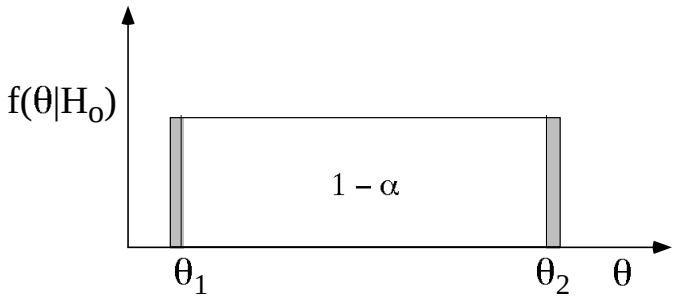

Figure 1.6: Would you accept this scheme to test hypotheses?

with the random variable $\overline{X}_{300}$ normally distributed because of the central limit theorem. This means that there is 99% probability that an average will come out in the interval $0.5 \pm (2.6 \times 0.017)$:

$$P(0.456 \leq \overline{X}_{300} \leq 0.544) = 99\% \,. \tag{1.10}$$

Imagine that a student obtains an average outside the above interval (e.g. $\overline{x} = 0.550$). The teacher may be interested in the probability that the student has made a mistake (for example, he has to decide if it is worthwhile checking the calculation in detail). Applying the standard methods one draws the conclusion that

> *"the hypothesis $H_\circ$ = 'no mistakes' is rejected at the 1% level of significance",*

i.e. one receives a precise answer to a different question. In fact, the meaning of the previous statement is simply

> *"there is only a 1% probability that the average falls outside the selected interval, if the calculations were done correctly".*

But this does not answer our natural question,[12] i.e. that concerning the probability of mistake, and not that of results far from the average if there were no mistakes. Moreover, the statement sounds as if one would be 99% sure that the student has made a mistake! This conclusion is highly misleading.

How is it possible, then, to answer the very question concerning the probability of mistakes? If you ask the students (before they take a standard course in hypothesis tests) you will hear the right answer, and it contains a crucial ingredient extraneous to the logic of hypothesis tests:

> *"It all depends on who has made the calculation!"*

In fact, if the calculation was done by a well-tested program the probability of mistake would be zero. And students know rather well their probability of making mistakes.

Example 6: A bad joke to a journal.[13]

[12] Personally, I find it is somehow impolite to give an answer to a question which is different from that asked. At least one should apologize for being unable to answer the original question. However, textbooks usually do not do this, and people get confused.

[13] Example taken from Ref. [12].

A scientific journal changes its publication policy. The editors announce that results with a significance level of 5% will no longer be accepted. Only those with a level of $\leq 1\%$ will be published. The rationale for the change, explained in an editorial, looks reasonable and it can be shared without hesitation: *"We want to publish only good results."*

1000 experimental physicists, not convinced by this severe rule, conspire against the journal. Each of them formulates a wrong physics hypothesis and performs an experiment to test it according to the accepted/rejected scheme.

Roughly 10 physicists get 1% significant results. Their papers are accepted and published. It follows that, contrary to the wishes of the editors, the first issue of the journal under the new policy contains only wrong results!

The solution to the kind of paradox raised by this example seems clear: The physicists knew with certainty that the hypotheses were wrong. So the example looks like an odd case with no practical importance. But in real life who knows in advance with certainty if a hypothesis is true or false?

1.9 Statistical significance versus probability of hypotheses

The examples in the previous section have shown the typical ways in which significance tests are misinterpreted. This kind of mistake is commonly made not only by students, but also by professional users of statistical methods. There are two different probabilities:

$P(H \mid \text{“}\mathbf{data}\text{”})$**:** the probability of the hypothesis H, conditioned by the observed data. This is the probabilistic statement in which we are interested. It summarizes the status of knowledge on H, achieved in conditions of uncertainty: it might be the probability that the W mass is between 80.00 and 80.50 GeV, that the Higgs mass is below 200 GeV, or that a charged track is a π^- rather than a K^-.

$P(\text{“}\mathbf{data}\text{”} \mid H)$**:** the probability of the observables under the condition that the hypothesis H is true.[14] For example, the probability of getting two consecutive heads when tossing a regular coin, the probability that a W mass is reconstructed within 1 GeV of the true mass, or that a 2.5 GeV pion produces a $\geq 100\,\mathrm{pC}$ signal in an electromagnetic calorimeter.

Unfortunately, conventional statistics considers only the second case. As a consequence, since the very question of interest remains unanswered, very often significance levels are incorrectly treated as if they were probabilities of the hypothesis. For example, *"H refused at 5% significance"* may be understood to mean the same as *"H has only 5% probability of being true."*

It is important to note the different consequences of the misunderstanding caused by the arbitrary probabilistic interpretation of confidence intervals and of significance levels. Measurement uncertainties on directly measured quantities obtained by confidence intervals are at least numerically correct in most routine cases, although arbitrarily interpreted. In hypothesis tests, however, the conclusions may become seriously wrong. This can be shown with the following examples.

Example 7: AIDS test.

An Italian citizen is chosen at random to undergo an AIDS test. Let us assume that the

[14] This should not be confused with the probability of the actual data, which is clearly 1, since they have been observed.

analysis used to test for HIV infection has the following performances:

$$P(Positive\,|\,HIV) \approx 1, \tag{1.11}$$
$$P(Positive\,|\,\overline{HIV}) = 0.2\%\,. \tag{1.12}$$

The analysis may declare healthy people 'Positive', even if only with a very small probability.

Let us assume that the analysis states 'Positive'. Can we say that, since the probability of an analysis error Healthy $\rightarrow$ Positive is only 0.2%, then the probability that the person is infected is 99.8%? Certainly not. If one calculates on the basis of an estimated 100 000 infected persons out of a population of 60 million, there is a 55% probability that the person is healthy![15] Some readers may be surprised to read that, in order to reach a conclusion, one needs to have an idea of how 'reasonable' the hypothesis is, independently of the data used: a mass cannot be negative; the spectrum of the true value is of a certain type; students often make mistakes; physical hypotheses happen to be incorrect; the proportion of Italians carrying the HIV virus is roughly 1 in 600. The notion of prior reasonableness of the hypothesis is fundamental to the approach we are going to present, but it is something to which physicists put up strong resistance (although in practice they often instinctively use this intuitive way of reasoning continuously and correctly). In this report I will try to show that 'priors' are rational and unavoidable, although their influence may become negligible when there is strong experimental evidence in favour of a given hypothesis.

Example 8: Probabilistic statements about the 1997 HERA high-Q^2 events.

A very instructive example of the misinterpretation of probability can be found in the statements which commented on the excess of events observed by the HERA experiments at DESY in the high-Q^2 region. For example, the official DESY statement [13] was:[16]

> *"The two HERA experiments, H1 and ZEUS, observe an excess of events above expectations at high x (or $M = \sqrt{x\,s}$), y, and Q^2. For $Q^2 > 15\,000$* GeV^2 the joint distribution has a probability of less than one per cent to come from Standard Model NC DIS processes."

Similar statements were spread out in the scientific community, and finally to the press. For example, a message circulated by INFN stated (it can be understood even in Italian)

"La probabilità che gli eventi osservati siano una fluttuazione statistica è inferiore all' 1%."

Obviously these two statements led the press (e.g. Corriere della Sera, 23 Feb. 1998) to

[15] The result will be a simple application of Bayes' theorem, which will be introduced later. A crude way to check this result is to imagine performing the test on the entire population. Then the number of persons declared Positive will be all the HIV infected plus 0.2% of the remaining population. In total 100 000 infected and 120 000 healthy persons. The general, Bayesian solution is given in Section 8.10.1

[16] One might think that the misleading meaning of that sentence was due to unfortunate wording, but this possibility is ruled out by other statements which show clearly a quite odd point of view of probabilistic matter. In fact the DESY 1998 activity report [14] insists that *"the likelihood that the data produced are the result of a statistical fluctuation ... is equivalent to that of tossing a coin and throwing seven heads or tails in a row"* (replacing 'probability' by 'likelihood' does not change the sense of the message). Then, trying to explain the meaning of a statistical fluctuation, the following example is given: *"This process can be simulated with a die. If the number of times a die is thrown is sufficiently large, the die falls equally often on all faces, i.e. all six numbers occur equally often. The probability for each face is exactly a sixth or 16.66%, assuming the die is not loaded. If the die is thrown less often, then the probability curve for the distribution of the six die values is no longer a straight line but has peaks and troughs. The probability distribution obtained by throwing the die varies about the theoretical value of 16.66% depending on how many times it is thrown."*

announce that scientists were highly confident that a great discovery was just around the corner.[17]

The experiments, on the other hand, did not mention this probability. Their published results [15] can be summarized, more or less, as *"there is a* $\lessapprox 1\%$ *probability of observing such events or rarer ones within the Standard Model".*

To sketch the flow of consecutive statements, let us indicate by SM *"the Standard Model is the only cause which can produce these events"* and by tail the *"possible observations which are rarer than the configuration of data actually observed".*

1. Experimental result: $P(data + tail \,|\, SM) \lesssim 1\%$.
2. Official statements: $P(SM \,|\, data) \lesssim 1\%$.
3. Press: $P(\overline{SM} \,|\, data) \gtrsim 99\%$, simply applying standard logic to the outcome of step 2. They deduce, correctly, that the hypothesis $\overline{SM}$ (= hint of new physics) is almost certain.

One can recognize an arbitrary inversion of probability. But now there is also something else, which is more subtle, and suspicious: *"why should we also take into account data which have not been observed?"*[18] Stated in a schematic way, it seems natural to draw conclusions on the basis of the observed data:

$$\textbf{data} \longrightarrow P(H \,|\, data)\,,$$

although $P(H \,|\, data)$ differs from $P(data \,|\, H)$. But it appears strange that unobserved data too should play a role. Nevertheless, because of our educational background, we are so used to the inferential scheme of the kind

$$\textbf{data} \longrightarrow P(H \,|\, data + tail)\,,$$

that we even have difficulty in understanding the meaning of this objection.[19]

Let us consider a new case, conceptually very similar, but easier to understand intuitively.

Example 9: Probability that a particular random number comes from a generator.
The value $x = 3.01$ is extracted from a Gaussian random-number generator having $\mu = 0$ and $\sigma = 1$. It is well known that

$$P(|X| > 3) = 0.27\%\,,$$

[17] One of the odd claims related to these events was on a poster of an INFN exhibition at Palazzo delle Esposizioni in Rome: *"These events are absolutely impossible within the current theory ... If they will be confirmed, it will imply that ... ."* Some friends of mine who visited the exhibition asked me what it meant that "something impossible needs to be confirmed".

[18] This is as if the conclusion from the AIDS test depended not only on $P(Positive \,|\, \overline{HIV})$ and on the prior probability of being infected, but also on the probability that this poor guy experienced events rarer than a mistaken analysis, like sitting next to Claudia Schiffer on an international flight, or winning the lottery, or being hit by a meteorite.

[19] I must admit I have fully understood this point only very recently, and I thank F. James for having asked, at the end of the CERN lectures, if I agreed with the sentence *"The probability of data not observed is irrelevant in making inferences from an experiment."* [10] I was not really ready to give a convincing reply, apart from a few intuitions, and from the trivial comment that this does not mean that we are not allowed to use MC data (strictly speaking, frequentists should not use MC data, as discussed in Section 8.1). In fact, in the lectures I did not talk about 'data+tails', but only about 'data'. This topic will be discussed again in Section 8.8.

but we cannot state that the value x has 0.27% probability of coming from that generator, or that the probability that the observation is a statistical fluctuation is 0.27%. In this case, the value comes with 100% probability from that generator, and it is at 100% a statistical fluctuation. This example helps to illustrate the logical mistake one can make in the previous examples. One may speak about the probability of the generator (let us call it A) only if another generator B is taken into account. If this is the case, the probability depends on the parameters of the generators, the observed value x and on the probability that the two generators enter the game. For example, if B has $\mu = 6.02$ and $\sigma = 1$, it is reasonable to think that

$$P(A \,|\, x = 3.01) = P(B \,|\, x = 3.01) = 0.5\,. \tag{1.13}$$

Let us imagine a variation of the example: The generation is performed according to an algorithm that chooses A or B, with a ratio of probability 10 to 1 in favour of A. The conclusions change: Given the same observed value $x = 3.01$, one would tend to infer that x is most probably due to A. It is not difficult to be convinced that, even if the value is a bit closer to the centre of generator B (for example $x = 3.3$), there will still be a tendency to attribute it to A. This natural way of reasoning is exactly what is meant by 'Bayesian', and will be illustrated in these notes.[20]. It should be noted that we are only considering the observed data ($x = 3.01$ or $x = 3.3$), and not other values which could be observed ($x \geq 3.01$, for example)

I hope these examples might at least persuade the reader to take the question of principles in probability statements seriously. Anyhow, even if we ignore philosophical aspects, there are other kinds of more technical inconsistencies in the way the standard paradigm is used to test hypotheses. These problems, which deserve extensive discussion, are effectively described in an interesting American Scientist article [10].

At this point I imagine that the reader will have a very spontaneous and legitimate objection: *"but why does this scheme of hypothesis tests usually work?"*. I will comment on this question in Section 8.8, but first we must introduce the alternative scheme for quantifying uncertainty.

[20] As an exercise, to compare the intuitive result with what we will learn later, it may be interesting to try to calculate, in the second case of the previous example ($P(A)/P(B) = 10$), the value x such that we would be in a condition of indifference (i.e. probability 50% each) with respect to the two generators.

Chapter 2

A probabilistic theory of measurement uncertainty

"If we were not ignorant there would be no probability, there could only be certainty. But our ignorance cannot be absolute, for then there would be no longer any probability at all. Thus the problems of probability may be classed according to the greater or less depth of our ignorance."
(Henri Poincaré)

2.1 Where to restart from?

In the light of the criticisms made in the previous chapter, it seems clear that we would be advised to completely revise the process which allows us to learn from experimental data. Paraphrasing Kant [16], one could say that (substituting the words in *italics* with those in parentheses):

> *"All metaphysicians (physicists) are therefore solemnly and legally suspended from their occupations till they shall have answered in a satisfactory manner the question, how are synthetic cognitions a priori possible (is it possible to learn from observations)?"*

Clearly this quotation must be taken in a playful way (at least as far as the invitation to suspended activities is concerned . . .). But, joking apart, the quotation is indeed more pertinent than one might initially think. In fact, Hume's criticism of the problem of induction, which interrupted the 'dogmatic slumber' of the great German philosopher, has survived the subsequent centuries.[1] We shall come back to this matter in a while.

[1] For example, it is interesting to report Einstein's opinion [17] about Hume's criticism: *"Hume saw clearly that certain concepts, as for example that of causality, cannot be deduced from the material of experience by logical methods. Kant, thoroughly convinced of the indispensability of certain concepts, took them – just as they are selected – to be necessary premises of every kind of thinking and differentiated them from concepts of empirical origin. I am convinced, however, that this differentiation is erroneous."* In the same Autobiographical Notes [17] Einstein, explaining how he came to the idea of the arbitrary character of absolute time, acknowledges that *"The type of critical reasoning which was required for the discovery of this central point was decisively furthered, in my case, especially by the reading of David Hume's and Ernst Mach's philosophical writings."* This tribute to Mach and Hume is repeated in the 'gemeinverständlich' of special relativity [18]: *"Why is it necessary to drag down from the Olympian fields of Plato the fundamental ideas of thought in natural science, and to attempt to reveal their earthly lineage? Answer: In order to free these ideas from the taboo attached to them, and thus to achieve greater freedom in the formation of ideas or concepts. It is to the immortal credit of D. Hume and E. Mach that they, above all others, introduced this critical conception."* I would like to end this parenthesis dedicated to Hume with a last citation, this time by de Finetti [11], closer to the argument of this chapter: *"In the philosophical*

In order to build a theory of measurement uncertainty which does not suffer from the problems illustrated above, we need to ground it on some kind of first principles, and derive the rest by logic. Otherwise we replace a collection of formulae and procedures handed down by tradition with another collection of cooking recipes.

We can start from two considerations.

1. In a way which is analogous to Descartes' *cogito*, the only statement with which it is difficult not to agree — in some sense the only certainty — is that (see end of Section 1.1)

   > *"the process of induction from experimental observations to statements about physics quantities (and, in general, physical hypothesis) is affected, unavoidably, by a certain degree of uncertainty".*

2. The natural concept developed by the human mind to quantify the plausibility of the statements in situations of uncertainty is that of probability.[2]

In other words we need to build a probabilistic (probabilistic and not, generically, statistic) theory of measurement uncertainty.

These two starting points seem perfectly reasonable, although the second appears to contradict the criticisms of the probabilistic interpretation of the result, raised above. However this is not really a problem, it is only a product of a distorted (i.e. different from the natural) view of the concept of probability. So, first we have to review the concept of probability. Once we have clarified this point, all the applications in measurement uncertainty will follow and there will be no need to inject *ad hoc* methods or use magic formulae, supported by authority but not by logic.

2.2 Concepts of probability

We have arrived at the point where it is necessary to define better what probability is. This is done in Section 3.3. As a general comment on the different approaches to probability, I would like, following Ref. [19], to cite de Finetti [11]:

> *"The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence. The numerous, different opposed attempts to put forward particular points of view which, in the opinion of their supporters, would endow Probability Theory with a 'nobler status', or a 'more scientific' character, or 'firmer' philosophical or logical foundations, have only served to generate confusion and obscurity, and to provoke well-known polemics and disagreements - even between supporters of essentially the same framework.*
>
> *The main points of view that have been put forward are as follows.*

arena, the problem of induction, its meaning, use and justification, has given rise to endless controversy, which, in the absence of an appropriate probabilistic framework, has inevitably been fruitless, leaving the major issues unresolved. It seems to me that the question was correctly formulated by Hume ... and the pragmatists ... However, the forces of reaction are always poised, armed with religious zeal, to defend holy obtuseness against the possibility of intelligent clarification. No sooner had Hume begun to prise apart the traditional edifice, then came poor Kant in a desperate attempt to paper over the cracks and contain the inductive argument — like its deductive counterpart — firmly within the narrow confines of the logic of certainty."

[2] Perhaps one may try to use instead fuzzy logic or something similar. I will only try to show that this way is productive and leads to a consistent theory of uncertainty which does not need continuous injections of extraneous matter. I am not interested in demonstrating the uniqueness of this solution, and all contributions on the subject are welcome.

> *The classical view is based on physical considerations of symmetry, in which one should be obliged to give the same probability to such 'symmetric' cases. But which 'symmetry'? And, in any case, why? The original sentence becomes meaningful if reversed: the symmetry is probabilistically significant, in someone's opinion, if it leads him to assign the same probabilities to such events.*
>
> *The logical view is similar, but much more superficial and irresponsible inasmuch as it is based on similarities or symmetries which no longer derive from the facts and their actual properties, but merely from sentences which describe them, and their formal structure or language.*
>
> *The frequentistic (or statistical) view presupposes that one accepts the classical view, in that it considers an event as a class of individual events, the latter being 'trials' of the former. The individual events not only have to be 'equally probable', but also 'stochastically independent' ... (these notions when applied to individual events are virtually impossible to define or explain in terms of the frequentistic interpretation). In this case, also, it is straightforward, by means of the subjective approach, to obtain, under the appropriate conditions, in perfectly valid manner, the result aimed at (but unattainable) in the statistical formulation. It suffices to make use of the notion of exchangeability. The result, which acts as a bridge connecting the new approach to the old, has often been referred to by the objectivists as "de Finetti's representation theorem'.*
>
> *It follows that all the three proposed definitions of 'objective' probability, although useless per se, turn out to be useful and good as valid auxiliary devices when included as such in the subjectivist theory."*

Also interesting is Hume's point of view on probability, where concept and evaluations are neatly separated. Note that these words were written in the middle of the 18th century [20].

> *"Though there be no such thing as Chance in the world; our ignorance of the real cause of any event has the same influence on the understanding, and begets a like species of belief or opinion.*
>
> *There is certainly a probability, which arises from a superiority of chances on any side; and according as this superiority increases, and surpasses the opposite chances, the probability receives a proportionable increase, and begets still a higher degree of belief or assent to that side, in which we discover the superiority. If a dye were marked with one figure or number of spots on four sides, and with another figure or number of spots on the two remaining sides, it would be more probable, that the former would turn up than the latter; though, if it had a thousand sides marked in the same manner, and only one side different, the probability would be much higher, and our belief or expectation of the event more steady and secure. This process of the thought or reasoning may seem trivial and obvious; but to those who consider it more narrowly, it may, perhaps, afford matter for curious speculation.*
>
> ...
>
> *Being determined by custom to transfer the past to the future, in all our inferences; where the past has been entirely regular and uniform, we expect the event with the greatest assurance, and leave no room for any contrary supposition. But where different effects have been found to follow from causes, which are to appearance exactly similar, all these various effects must occur to the mind in transferring the past to the future, and enter into our consideration, when we determine the probability of the event. Though we give the preference to that which has been found most usual, and believe that this effect will exist, we must not overlook the other effects, but must assign to each of them a particular weight and authority, in proportion as we have found it to be more or less frequent."*

2.3 Subjective probability

I would like to sketch the essential concepts related to subjective probability,[3] for the convenience of those who wish to have a short overview of the subject, discussed in detail in Part II. This should also help those who are not familiar with this approach to follow the scheme of probabilistic induction which will be presented in the next section, and the summary of the applications which will be developed in the rest of the notes.

- Essentially, one assumes that the concept of probability is primitive, i.e. close to that of common sense (said with a joke, probability is what everybody knows before going to school and continues to use afterwards, in spite of what one has been taught[4]).
- Stated in other words, probability is a measure of the degree of belief that any well-defined proposition (an event) will turn out to be true.
- Probability is related to the state of uncertainty, and not (only) to the outcome of repeated experiments.
- The value of probability ranges between 0 and 1 from events which go from false to true (see Fig. 3.1 in Section 3.3.2).
- Since the more one believes in an event the more money one is prepared to bet, the 'coherent' bet can be used to define the value of probability in an operational way (see Section 3.3.2).
- From the condition of coherence one obtains, as theorems, the basic rules of probability (usually known as axioms) and the 'formula of conditional probability' (see Sections 3.4.2 and 8.3).
- There is, in principle, an infinite number of ways to evaluate the probability, with the only condition being that they must satisfy coherence. We can use symmetry arguments, statistical data (past frequencies), Monte Carlo simulations, quantum mechanics[5] and so on. What is important is that if we get a number close to one, we are very confident that the event will happen; if the number is close to zero we are very confident that it will not happen; if $P(A) > P(B)$, then we believe in the realization of A more than in the realization of B.
- It is easy to show that the usual 'definitions' suffer from circularity[6] (Section 3.3.1), and that they can be used only in very simple and stereotypical cases. In the subjective approach they can be easily recovered as 'evaluation rules' under appropriate conditions.

[3] For an introductory and concise presentation of the subject see also Ref. [21].

[4] This remark — not completely a joke — is due to the observation that most physicists interviewed are convinced that (1.3) is legitimate, although they maintain that probability is the limit of the frequency.

[5] Without entering into the open problems of quantum mechanics, let us just say that it does not matter, from the cognitive point of view, whether one believes that the fundamental laws are intrinsically probabilistic, or whether this is just due to a limitation of our knowledge, as hidden variables *à la Einstein* would imply. If we calculate that process A has a probability of 0.9, and process B 0.4, we will believe A much more than B.

[6] Concerning the combinatorial definition, Poincaré's criticism [6] is remarkable:

> *"The definition, it will be said, is very simple. The probability of an event is the ratio of the number of cases favourable to the event to the total number of possible cases. A simple example will show how incomplete this definition is: ...*
> *... We are therefore bound to complete the definition by saying '... to the total number of possible cases, provided the cases are equally probable.' So we are compelled to define the probable by the probable. How can we know that two possible cases are equally probable? Will it be by convention? If we insert at the beginning of every problem an explicit convention, well and good! We then have*

- Subjective probability becomes the most general framework, which is valid in all practical situations and, particularly, in treating uncertainty in measurements.

- Subjective probability does not mean arbitrary[7]; on the contrary, since the normative role of coherence morally obliges a person who assesses a probability to take personal responsibility, he will try to act in the most objective way (as perceived by common sense).

- The word 'belief' can hurt those who think, naïvely, that in science there is no place for beliefs. This point will be discussed in more detail in Section 8.4. For an extensive discussion see Ref. [22].

- Objectivity is recovered if rational individuals share the same culture and the same knowledge about experimental data, as happens for most textbook physics; but one should speak, more appropriately, of intersubjectivity.

- The utility of subjective probability in measurement uncertainty has already been recognized[8] by the aforementioned ISO Guide [3], after many internal discussions (see Ref. [23] and references therein):

  > *"In contrast to this frequency-based point of view of probability an equally valid viewpoint is that probability is a measure of the degree of belief that an event will occur ... Recommendation INC-1 ... implicitly adopts such a viewpoint of probability."*

- In the subjective approach random variables (or, better, uncertain numbers) assume a more general meaning than that they have in the frequentistic approach: a random number is just any number in respect of which one is in a condition of uncertainty. For example:

  1. if I put a reference weight (1 kg) on a balance with digital indication to the centigramme, then the random variable is the value (in grammes) that I am expected to read (X): 1000.00, 999.95 ... 1000.03 ... ?
  2. if I put a weight of unknown value and I read 576.23 g, then the random value (in grammes) becomes the mass of the body (μ): 576.10, 576.12 ... 576.23 ... 576.50 ... ?

  In the first case the random number is linked to observations, in the second to true values.

- The different values of the random variable are classified by a function $f(x)$ which quantifies the degree of belief of all the possible values of the quantity.

---

nothing to do but to apply the rules of arithmetic and algebra, and we complete our calculation, when our result cannot be called in question. But if we wish to make the slightest application of this result, we must prove that our convention is legitimate, and we shall find ourselves in the presence of the very difficulty we thought we had avoided."

[7] Perhaps this is the reason why Poincaré [6], despite his many brilliant intuitions, above all about the necessity of the priors (*"there are certain points which seem to be well established. To undertake the calculation of any probability, and even for that calculation to have any meaning at all, we must admit, as a point of departure, an hypothesis or convention which has always something arbitrary on it ...*), concludes to *"... have set several problems, and have given no solution ..."*. The coherence makes the distinction between arbitrariness and 'subjectivity' and gives a real sense to subjective probability.

[8] One should feel obliged to follow this recommendation as a metrology rule. It is however remarkable to hear that, in spite of the diffused cultural prejudices against subjective probability, the scientists of the ISO working groups have arrived at such a conclusion.

- All the formal properties of $f(x)$ are the same as in conventional statistics (average, variance, etc.).

- All probability distributions are conditioned to a given state of information: in the examples of the balance one should write, more correctly,

$$\begin{aligned} f(x) &\longrightarrow f(x \,|\, \mu = 1000.00) \\ f(\mu) &\longrightarrow f(\mu \,|\, x = 576.23)\,. \end{aligned}$$

- Of particular interest is the special meaning of conditional probability within the framework of subjective probability. Also in this case this concept turns out to be very natural, and the subjective point of view solves some paradoxes of the so-called 'definition' of conditional probability (see Section 8.3).

- The subjective approach is often called Bayesian, because of the central role of Bayes' theorem, which will be introduced in Section 2.6. However, although Bayes' theorem is important, especially in scientific applications, one should not think that this is the only way to assess probabilities. Outside the well-specified conditions in which it is valid, the only guidance is that of coherence.

- Considering the result of a measurement, the entire state of uncertainty is held in $f(\mu)$; then one may calculate intervals in which we think there is a given probability to find μ, value(s) of maximum belief (mode), average, standard deviation, etc., which allow the result to be summarized with only a couple of numbers, chosen in a conventional way.

2.4 Learning from observations: the 'problem of induction'

Having briefly shown the language for treating uncertainty in a probabilistic way, it remains now to see how one builds the function $f(\mu)$ which describes the beliefs in the different possible values of the physics quantity. Before presenting the formal framework we still need a short introduction on the link between observations and hypotheses.

Every measurement is made with the purpose of increasing the knowledge of the person who performs it, and of anybody else who may be interested in it. This may be the members of a scientific community, a physician who has prescribed a certain analysis or a merchant who wants to buy a certain product. It is clear that the need to perform a measurement indicates that one is in a state of uncertainty with respect to something, e.g. a fundamental constant of physics or a theory of the Universe; the state of health of a patient; the chemical composition of a product. In all cases, the measurement has the purpose of modifying a given state of knowledge. One would be tempted to say 'acquire', instead of 'modify', the state of knowledge, thus indicating that the knowledge could be created from nothing with the act of the measurement. Instead, it is not difficult to realize that, in all cases, it is just an updating process, in the light of new facts and of some reason. Let us take the example of the measurement of the temperature in a room, using a digital thermometer — just to avoid uncertainties in the reading — and let us suppose that we get 21.7 °C. Although we may be uncertain on the tenths of a degree, there is no doubt that the measurement will have squeezed the interval of temperatures considered to be possible before the measurement: those compatible with the physiological feeling of 'comfortable environment'. According to our knowledge of the thermometer used, or of thermometers in general, there will

be values of temperature in a given interval around 21.7 °C which we believe more and values outside which we believe less.[9]

It is, however, also clear that if the thermometer had indicated, for the same physiological feeling, 17.3 °C, we might think that it was not well calibrated. There would be, however, no doubt that the instrument was not working properly if it had indicated 2.5 °C!

The three cases correspond to three different degrees of modification of the knowledge. In particular, in the last case the modification is null.[10]

The process of learning from empirical observations is called induction by philosophers. Most readers will be aware that in philosophy there exists the unsolved 'problem of induction', raised by Hume. His criticism can be summarized by simply saying that induction is not justified, in the sense that observations do not lead necessarily (with the logical strength of a mathematical theorem) to certain conclusions. The probabilistic approach adopted here seems to be the only reasonable way out of such a criticism.

2.5 Beyond Popper's falsification scheme

People very often think that the only scientific method valid in physics is that of Popper's falsification scheme. There is no doubt that, if a theory is not capable of explaining experimental results, it should be rejected or modified. But, since it is impossible to demonstrate with certainty that a theory is true, it becomes impossible to decide among the infinite number of hypotheses which have not been falsified. This would produce stagnation in research. A probabilistic method allows, instead, for a scale of credibility to be provided for classifying all hypotheses taken into account (or credibility ratios between any pair of hypotheses). This is close to the natural development of science, where new investigations are made in the direction which seems the most credible, according to the state of knowledge at the moment at which the decision on how to proceed was made.

As far as the results of measurements are concerned, the falsification scheme is absolutely unsuitable. Taking it literally, one should be authorized only to check whether or not the value read on an instrument is compatible with a true value, nothing more. It is understandable then that, with this premise, one cannot go very far.

We will show that falsification is just a subcase of the Bayesian inference.

2.6 From the probability of the effects to the probability of the causes

The scheme of updating knowledge that we will use is that of Bayesian statistical inference, widely discussed in the second part of this report (in particular Sections 3.4 and 5.2). I wish to make a less formal presentation of it here, to show that there is nothing mysterious behind Bayes' theorem, and I will try to justify it in a simple way.

It is very convenient to consider true values and observed values as causes and effects (see Fig. 2.1, imagining also a continuous set of causes and many possible effects). The process of going from causes to effects it is called 'deduction'.[11] The possible values x which may be

[9] To understand the role of implicit prior knowledge, imagine someone having no scientific or technical education at all, entering a physics laboratory and reading a number on an instrument. His scientific knowledge will not improve at all, apart from the triviality that a given instrument displayed a number (not much knowledge).

[10] But also in this case we have learned something: the thermometer does not work.

[11] To be correct, the deduction we are talking about is different from the classical one. We are dealing, in fact, with probabilistic deduction, in the sense that, given a certain cause, the effect is not univocally determined.

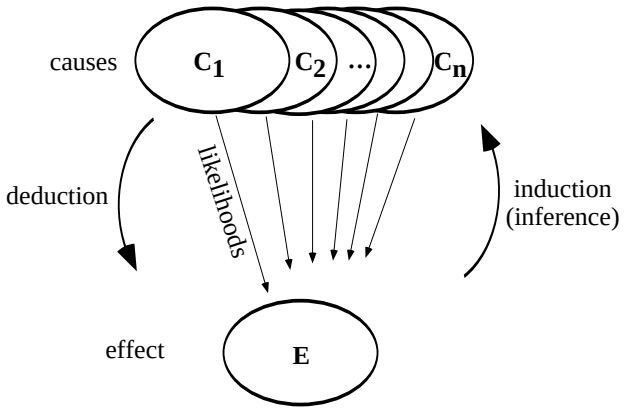

Figure 2.1: Deduction and induction.

observed are classified in belief by

$$f(x \,|\, \mu)\,.$$

This function is called 'likelihood' since it quantifies how likely it is that μ will produce any given x. It summarizes all previous knowledge on that kind of measurement (behaviour of the instruments, of influence factors, etc. – see list in Section 1.3). Often, if one deals only with random error, the $f(x \,|\, \mu)$ is a normal distribution around μ, but in principle it may have any form.

Once the likelihood is determined (we have the performance of the detector under control) we can build $f(\mu \,|\, x)$, under the hypothesis that x will be observed.[12] In order to arrive at the general formula in an heuristic way, let us consider only two values of μ. If they seem to us equally possible, it will seem natural to be in favour of the value which gives the highest likelihood that x will be observed. For example, assuming $\mu_1 = -1$, $\mu_2 = 10$, considering a normal likelihood with $\sigma = 3$, and having observed $x = 2$, one tends to believe that the observation is most likely caused by μ_1. If, on the other hand, the quantity of interest is positively defined, then μ_1 switches from most probable to impossible cause; μ_2 becomes certain. There are, in general, intermediate cases in which, because of previous knowledge (see, e.g., Fig. 1.3 and related text), one tends to believe *a priori* more in one or other of the causes. It follows that, in the light of a new observation, the degree of belief of a given value of μ will be proportional to

- the likelihood that μ will produce the observed effect;
- the degree of belief attributed to μ before the observation, quantified by $f_\circ(\mu)$.

We have finally:

$$f(\mu \,|\, x) \propto f(x \,|\, \mu) \cdot f_\circ(\mu)\,.$$

This is one of the ways to write Bayes' theorem.

[12] It is important to understand that $f(\mu \,|\, x)$ can be evaluated before one knows the observed value x. In fact, to be correct, $f(\mu \,|\, x)$ should be interpreted as beliefs of μ under the hypothesis that x is observed, and not only as beliefs of μ after x is observed. Similarly, $f(x \,|\, \mu)$ can also be built after the data have been observed, although for teaching purposes the opposite has been suggested, which corresponds to the most common case.

2.7 Bayes' theorem for uncertain quantities: derivation from a physicist's point of view

Let us show a little more formally the concepts illustrated in the previous section. This is proof of the Bayes' theorem alternative to the proof applied to events, given in Part II of these notes. It is now applied directly to uncertain (i.e. random) quantities, and it should be closer to the physicist's reasoning than the standard proof. For teaching purposes I explain it using time ordering, but this is unnecessary, as explained several times elsewhere.

- Before doing the experiment we are uncertain of the values of μ and x: we know neither the true value, nor the observed value. Generally speaking, this uncertainty is quantified by $f(x, \mu)$.
- Under the hypothesis that we observe x, we can calculate the conditional probability
$$f(\mu \,|\, x) = \frac{f(x, \mu)}{f(x)} = \frac{f(x, \mu)}{\int f(x, \mu)\, \mathrm{d}\mu}\,.$$
- Usually we don't have $f(x, \mu)$, but this can be calculated by $f(x \,|\, \mu)$ and $f(\mu)$:
$$f(x, \mu) = f(x \,|\, \mu) \cdot f(\mu)\,.$$
- If we do an experiment we need to have a good idea of the behaviour of the apparatus; therefore $f(x \,|\, \mu)$ must be a narrow distribution, and the most imprecise factor remains the knowledge about μ, quantified by $f(\mu)$, usually very broad. But it is all right that this should be so, because we want to learn about μ.
- Putting all the pieces together we get the standard formula of Bayes' theorem for uncertain quantities:
$$f(\mu \,|\, x) = \frac{f(x \,|\, \mu) \cdot f(\mu)}{\int f(x \,|\, \mu) \cdot f(\mu)\, \mathrm{d}\mu}\,.$$

The steps followed in this proof of the theorem should convince the reader that $f(\mu \,|\, x)$ calculated in this way is the best we can say about μ with the given status of information.

2.8 Afraid of 'prejudices'? Inevitability of principle and frequent practical irrelevance of the priors

Doubtless, many readers could be at a loss at having to accept that scientific conclusions may depend on prejudices about the value of a physical quantity ('prejudice' currently has a negative meaning, but in reality it simply means 'scientific judgement based on previous experience'). We shall have many opportunities to enter again into discussion about this problem, but it is important to give a general overview now and to make some firm statements on the role of priors.

- First, from a theoretical point of view, it is impossible to get rid of priors; that is if we want to calculate the probability of events of practical interest, and not just solve mathematical games.
- At a more intuitive level, it is absolutely reasonable to draw conclusions in the light of some reason, rather than in a purely automatic way.

- In routine measurements the interval of prior acceptance of the possible values is so large, compared to the width of the likelihood (seen as a function of μ), that, in practice, it is as if all values were equally possible. The prior is then absorbed into the normalization constant:

$$f(x \,|\, \mu) \cdot f_\circ(\mu) \xrightarrow[\textit{prior very vague}]{} f(x \,|\, \mu). \tag{2.1}$$

- If, instead, this is not the case, it is absolutely legitimate to believe more in personal prejudices than in empirical data. This could be when one uses an instrument of which one is not very confident, or when one does for the first time measurements in a new field, or in a new kinematical domain, and so on. For example, it is easier to believe that a student has made a trivial mistake than to conceive that he has discovered a new physical effect. An interesting case is mentioned by Poincaré [6]:

> *"The impossibility of squaring the circle was shown in 1885, but before that date all geometers considered this impossibility as so 'probable' that the Académie des Sciences rejected without examination the, alas! too numerous memoirs on this subject that a few unhappy madmen sent in every year. Was the Académie wrong? Evidently not, and it knew perfectly well that by acting in this manner it did not run the least risk of stifling a discovery of moment. The Académie could not have proved that it was right, but it knew quite well that its instinct did not deceive it. If you had asked the Academicians, they would have answered: 'We have compared the probability that an unknown scientist should have found out what has been vainly sought for so long, with the probability that there is one madman the more on the earth, and the latter has appeared to us the greater.'"*

In conclusion, contrary to those who try to find objective priors which would give the Bayesian theory a nobler status of objectivity, I prefer to state explicitly the naturalness and necessity of subjective priors [22]. If rational people (e.g. physicists), under the guidance of coherency (i.e. they are honest), but each with unavoidable personal experience, have priors which are so different that they reach divergent conclusions, it just means that the data are still not sufficiently solid to allow a high degree of intersubjectivity (i.e. the subject is still in the area of active research rather than in that of consolidated scientific culture). On the other hand, the step from abstract objective rules to dogmatism is very short [22].

Turning now to the more practical aspect of presenting a result, I will give some recommendations about unbiased ways of doing this, in cases when priors are really critical (Section 9.2). Nevertheless, it should be clear that:

- since the natural conclusions should be probabilistic statements on physical quantities, someone has to turn the likelihoods into probabilities, and those who have done the experiment are usually the best candidates for doing this;
- taking the spirit of publishing unbiased results — which is in principle respectable — to extremes, one should not publish any result, but just raw data tapes.

2.9 Recovering standard methods and short-cuts to Bayesian reasoning

Before moving on to applications, it is necessary to answer an important question: *"Should one proceed by applying Bayes' theorem in every situation?"* The answer is no, and the alternative

is essentially implicit in (2.1), and can be paraphrased with the example of the dog and the hunter.

We have already used this example in Section 1.7, when we were discussing the arbitrariness of probability inversion performed unconsciously by (most of)[13] those who use the scheme of confidence intervals. The same example will also be used in Section 5.2.3, when discussing the reason why Bayesian estimators appear to be distorted (a topic discussed in more detail in Section 8.5). This analogy is very important, and, in many practical applications, it allows us to bypass the explicit use of Bayes' theorem when priors do not sizably influence the result (in the case of a normal model the demonstration can be seen in Section 5.4.2).

Figure 2.2 shows how it is possible to recover standard methods from a Bayesian perspective. One sees that the crucial link is with the Maximum Likelihood Principle, which, in this approach is just a subcase (see Section 5.2.2). Then, when extra simplifying restrictions are verified, the different forms of the Least Squares are reobtained. In conclusion:

- One is allowed to use these methods if one thinks that the approximations are valid; the same happens with the usual propagation of uncertainties and of their correlations, outlined in the next section.

- One keeps the Bayesian interpretation of the results; in particular, one is allowed to talk about the probability distributions of the true values, with all the philosophical and practical advantages we have seen.

- Even if the priors are not negligible, but the final distribution is roughly normal,[14] one can evaluate the expected value and standard deviation from the shape of the distribution, as is well known:

$$\frac{\partial \ln f(\mu \,|\, x)}{\partial \mu} = 0 \quad \Rightarrow \quad E(\mu) \approx \mu_m, \tag{2.2}$$

$$-\frac{\partial^2 \ln f(\mu \,|\, x)}{\partial \mu^2}\bigg|_{\mu_m} \quad \Rightarrow \quad \approx \frac{1}{Var(\mu)}\,, \tag{2.3}$$

  where μ_m stands for the mode of the distribution.

2.10 Evaluation of uncertainty: general scheme

Now that we have set up the framework, we can draw the general scheme to evaluate uncertainty in measurement in the most general cases. For the basic applications we will refer to the "primer" and to the appendix. For more sophisticated applications the reader is recommended to search in specialized literature.

2.10.1 Direct measurement in the absence of systematic errors

The first step consists in evaluating the uncertainty on a quantity measured directly. The most common likelihoods which describe the observed values are the Gaussian, the binomial and the Poisson distributions.

[13] Although I don't believe it, I leave open the possibility that there really is someone who has developed some special reasoning to avoid, deep in his mind, the category of the probable when figuring out the uncertainty on a true value.

[14] In case of doubt it is recommended to plot it.

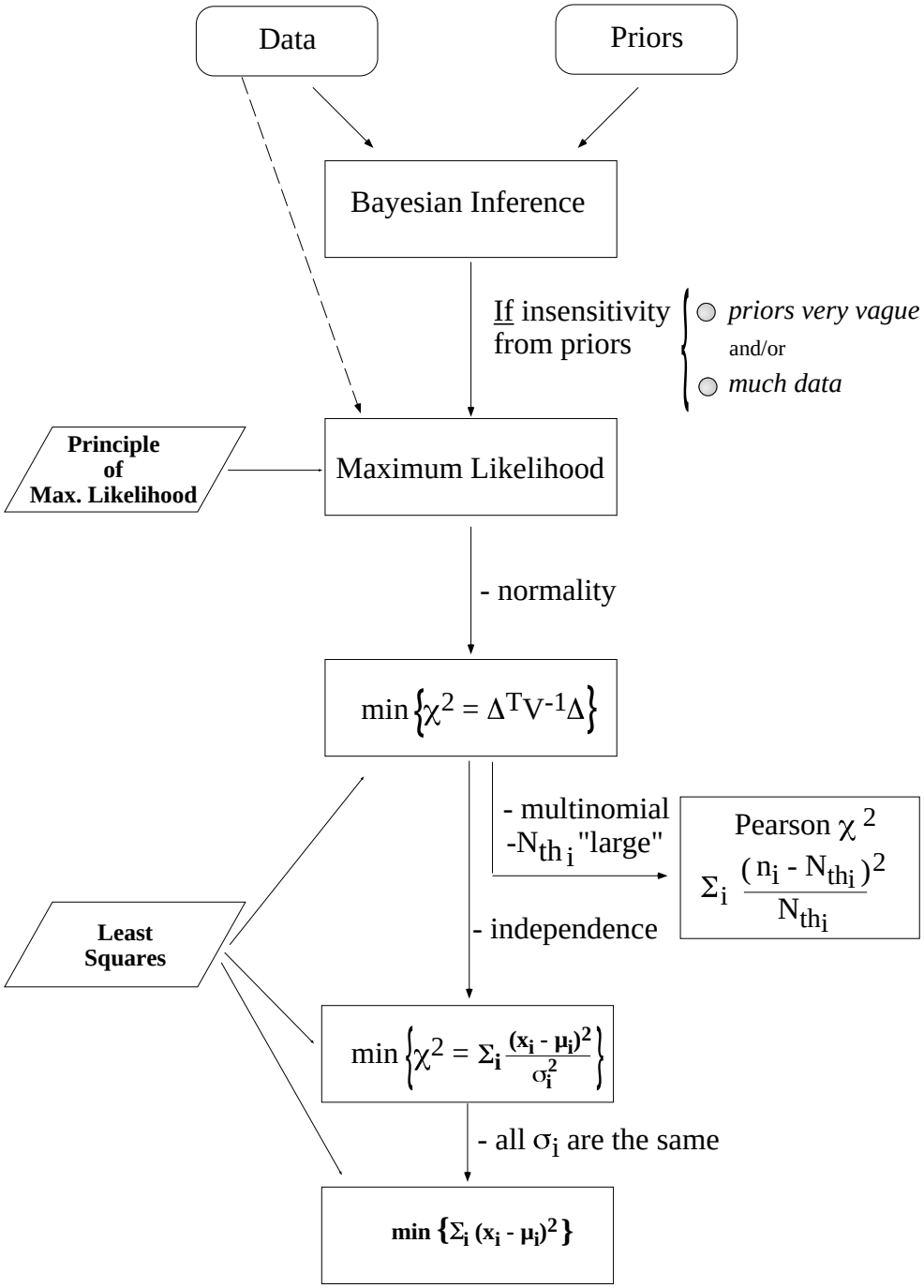

Figure 2.2: Relation between Bayesian inference and standard data analysis methods. The top-down flow shows subsequent limiting conditions. For an understanding of the relation between the 'normal' χ^2 and the Pearson χ^2 Ref. [24] is recommended.

Gaussian: This is the well-known case of 'normally' distributed errors. For simplicity, we will only consider σ independent of μ (constant r.m.s. error within the range of measurability), but there is no difficulty of principle in treating the general case. The following cases will be analysed:

- inference on μ starting from a prior much more vague than the width of the likelihood (Section 5.4.1);
- prior width comparable with that of the likelihood (Section 5.4.2): this case also describes the combination of independent measurements;
- observed values very close to, or beyond the edge of the physical region (Section 5.4.3);
- a method to give unbiased estimates will be discussed in Sections 9.2 and 9.2.1, but at the cost of having to introduce fictitious quantities.

Binomial: This distribution is important for efficiencies and, in the general case, for making inferences on unknown proportions. The cases considered include (see Section 5.5.1):

- general case with flat prior leading to the recursive Laplace formula (the problem solved originally by Bayes);
- limit to normality;
- combinations of different datasets coming from the same proportion;
- upper and lower limits when the efficiency is 0 or 1;
- comparison with Poisson approximation.

Poisson: The cases of counting experiments here considered[15] are (see Section 5.5.2):

- inference on λ starting from a flat distribution;
- upper limit in the case of null observation;
- counting measurements in the presence of a background, when its rate is well known (Sections 5.6.5 and 9.1.6);
- more complicated case of background with an uncertain rate (Section 5.6.5);
- dependence of the conclusions on the choice of experience-motivated priors (Section 9.1);
- combination of upper limits, also considering experiments of different sensitivity (Section 9.1.3).
- effect of possible systematic errors (Section 9.1.4);
- a special section will be dedicated to the lower bounds on the mass of a new hypothetical particle from counting experiments and from direct information (Section 9.3).

[15] For a general and self-contained discussion concerning the inference of the intensity of Poisson processes at the limit of the detector sensitivity, see Ref. [25].

2.10.2 Indirect measurements

The case of quantities measured indirectly is conceptually very easy, as there is nothing to 'think'. Since all values of the quantities are associated with random numbers, the uncertainty on the input quantities is propagated to that of output quantities, making use of the rules of probability. Calling μ_1, μ_2 and μ_3 the generic quantities, the inferential scheme is:

$$\left.\begin{array}{l} f(\mu_1 \,|\, data_1) \\ f(\mu_2 \,|\, data_2) \end{array}\right\} \xrightarrow[\mu_3 = g(\mu_1,\mu_2)]{} f(\mu_3 \,|\, data_1, data_2)\,. \tag{2.4}$$

The problem of going from the probability density functions (p.d.f.'s) of μ_1 and μ_2 to that of μ_3 makes use of probability calculus, which can become difficult, or impossible to do analytically, if p.d.f.'s or $g(\mu_1, \mu_2)$ are complicated mathematical functions. Anyhow, it is interesting to note that the solution to the problem is, indeed, simple, at least in principle. In fact, $f(\mu_3)$ is given, in the most general case, by

$$f(\mu_3) = \int f(\mu_1) \cdot f(\mu_2) \cdot \delta(y_3 - g(\mu_1, \mu_2))\, \mathrm{d}\mu_1 \mathrm{d}\mu_2\,, \tag{2.5}$$

where $\delta()$ is the Dirac delta function. The formula can be easily extended to many variables, or even correlations can be taken into account (one needs only to replace the product of individual p.d.f.'s by a joint p.d.f.). Equation (2.5) has a simple intuitive interpretation: the infinitesimal probability element $f(\mu_3)\, d\mu_3$ depends on 'how many' (we are dealing with infinities!) elements $d\mu_1 d\mu_2$ contribute to it, each weighed with the p.d.f. calculated in the point $\{\mu_1, \mu_2\}$. An alternative interpretation of Eq. (2.5), very useful in applications, is to think of a Monte Carlo simulation, where all possible values of μ_1 and μ_2 enter with their distributions, and correlations are properly taken into account. The histogram of μ_3 calculated from $\mu_3 = g(\mu_1, \mu_2)$ will 'tend' to $f(\mu_3)$ for a large number of generated events.[16]

In routine cases the propagation is done in an approximate way, assuming linearization of $g(\mu_1, \mu_2)$ and normal distribution of μ_3. Therefore only variances and covariances need to be calculated. The well-known error propagation formulae are recovered (Section 2.10.4), but now with a well-defined probabilistic meaning.

2.10.3 Systematic errors

Uncertainty due to systematic effects is also included in a natural way in this approach. Let us first define the notation (i is the generic index):

- $\underline{x} = \{x_1, x_2, \ldots x_{n_x}\}$ is the 'n-tuple' (vector) of observables X_i;
- $\underline{\mu} = \{\mu_1, \mu_2, \ldots \mu_{n_\mu}\}$ is the n-tuple of true values μ_i;
- $\underline{h} = \{h_1, h_2, \ldots h_{n_h}\}$ is the n-tuple of influence quantities H_i.

By influence quantities we mean:

→ all kinds of external factors which may influence the result (temperature, atmospheric pressure, etc.);

[16] As we shall see, the use of frequencies is absolutely legitimate in subjective probability, once the distinction between probability and frequency is properly made. In this case it works because of the Bernoulli theorem, which states that for a very large Monte Carlo sample *"it is very improbable that the frequency distribution will differ much from the p.d.f."* (This is the probabilistic meaning to be attributed to 'tend'.)

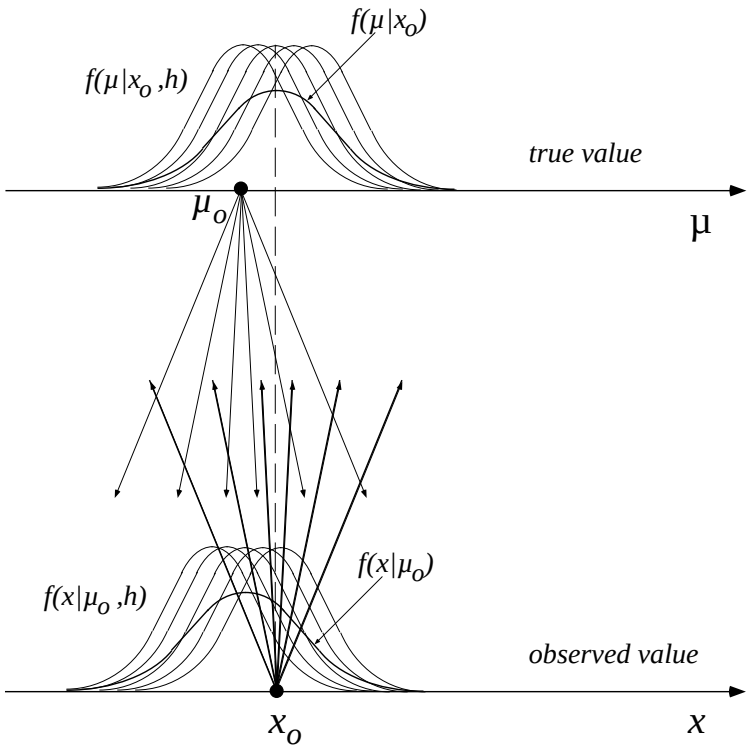

Figure 2.3: Model to handle the uncertainty due to systematic errors by the use of conditional probability.

→ all calibration constants;

→ all possible hypotheses upon which the results may depend (e.g. Monte Carlo parameters).

From a probabilistic point of view, there is no distinction between $\underline{\mu}$ and $\underline{h}$: they are all conditional hypotheses for the $\underline{x}$, i.e. causes which produce the observed effects. The difference is simply that we are interested in $\underline{\mu}$ rather than in $\underline{h}$.[17]

There are alternative ways to take into account the systematic effects in the final distribution of $\underline{\mu}$:

1. Global inference on $f(\underline{\mu}, \underline{h})$. We can use Bayes' theorem to make an inference on $\underline{\mu}$ and $\underline{h}$, as described in Section 5.2.1:

$$\underline{x} \Rightarrow f(\underline{\mu}, \underline{h} \,|\, \underline{x}) \Rightarrow f(\underline{\mu} \,|\, \underline{x})\,.$$

   This method, depending on the joint prior distribution $f_\circ(\underline{\mu}, \underline{h})$, can even model possible correlations between $\underline{\mu}$ and $\underline{h}$ (e.g. radiative correction depending on the quantity of interest).

2. Conditional inference (see Fig. 2.3). Given the observed data, one has a joint distribution

[17] For example, in the absence of random error the reading (X) of a voltmeter depends on the probed voltage (V) and on the scale offset (Z): $X = V - Z$. Therefore, the result from the observation of $X = x$ gives only a constraint between V and Z:

$$V - Z = x\,.$$

If we know Z well (within unavoidable uncertainty), then we can learn something about V. If instead the prior knowledge on V is better than that on Z we can use the measurement to calibrate the instrument.

of $\underline{\mu}$ for all possible configurations of $\underline{h}$:

$$\underline{x} \Rightarrow f(\underline{\mu} \,|\, \underline{x}, \underline{h}) \,.$$

Each conditional result is reweighed with the distribution of beliefs of $\underline{h}$, using the well-known law of probability:

$$f(\underline{\mu} \,|\, \underline{x}) = \int f(\underline{\mu} \,|\, \underline{x}, \underline{h}) \cdot f(\underline{h}) \, \mathrm{d}\underline{h} \,. \tag{2.6}$$

3. Propagation of uncertainties. Essentially, one applies the propagation of uncertainty, whose most general case has been illustrated in the previous section, making use of the following model: One considers a raw result on raw values $\underline{\mu}_R$ for some nominal values of the influence quantities, i.e.

$$f(\underline{\mu}_R \,|\, \underline{x}, \underline{h}_\circ) \,;$$

then (corrected) true values are obtained as a function of the raw ones and of the possible values of the influence quantities, i.e.

$$\mu_i = \mu_i(\mu_{i_R}, \underline{h}) \,.$$

The three ways lead to the same result and each of them can be more or less intuitive to different people, and more less suitable for different applications. For example, the last two, which are formally equivalent, are the most intuitive for HEP experimentalists, and it is conceptually equivalent to what they do when they vary — within reasonable intervals — all Monte Carlo parameters in order to estimate the systematic errors.[18] The third form is particularly convenient to make linear expansions which lead to approximated solutions (see Section 6.1).

There is an important remark to be made. In some cases it is preferable not to 'integrate'[19] over all h's. Instead, it is better to report the result as $f(\underline{\mu} \,|\, \{h\})$, where $\{h\}$ stands for a subset of $\underline{h}$, taken at their nominal values, if:

- $\{h\}$ could be controlled better by the users of the result (for example $h_i \in \{h\}$ is a theoretical quantity on which there is work in progress);
- there is some chance of achieving a better knowledge of $\{h\}$ within the same experiment (for example h_i could be the overall calibration constant of a calorimeter);
- a discrete and small number of very different hypotheses could affect the result. For example:

$$\begin{aligned} f(\alpha_s \,|\, Th_1, \mathcal{O}(\alpha_s^2), \dots) &= \dots \\ f(\alpha_s \,|\, Th_2, \mathcal{O}(\alpha_s^2), \dots) &= \dots \,. \end{aligned}$$

This is, in fact, the standard way in which this kind of result has been presented (apart from the inessential fact that only best values and standard deviations are given, assuming normality).

If results are presented under the condition of $\{h\}$, one should also report the derivatives of $\underline{\mu}$ with respect to the result, so that one does not have to redo the complete analysis when the influence factors are better known. A typical example in which this is usually done is the possible variation of the result due to the precise values of the charm-quark mass. A recent example in which this idea has been applied thoroughly is given in Ref. [26].

[18] But, in order to give a well-defined probabilistic meaning to the result, the variations must be performed according to $f(\underline{h})$, and not arbitrary.

[19] 'Integrate' stands for a generic term which also includes the approximate method just described.

2.10.4 Approximate methods

Of extreme practical importance are the approximate methods, which enable us not only to avoid having to use Bayes' theorem explicitly, but also to avoid working with probability distributions. In particular, propagation of uncertainty, including due to statistical effects of unknown size, is done in this way in all routine applications, as has been remarked in the previous section. These methods are discussed in Chapter 6, together with some words of caution about their uncritical use (see Sections 6.1.5, 6.2 and 6.3.2).

Part II

Bayesian primer

- slightly reviewed version of the 1995 DESY/Rome report -

Chapter 3

Subjective probability and Bayes' theorem

"The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence."
(Bruno de Finetti)

3.1 Original abstract of the primer

Bayesian statistics is based on the subjective definition of probability as 'degree of belief' and on Bayes' theorem, the basic tool for assigning probabilities to hypotheses combining *a priori* judgements and experimental information. This was the original point of view of Bayes, Bernoulli, Gauss, Laplace, etc. and contrasts with later conventional (pseudo-)definitions of probabilities, which implicitly presuppose the concept of probability. These notes[1] show that the Bayesian approach is the natural one for data analysis in the most general sense, and for assigning uncertainties to the results of physical measurements, while at the same time resolving philosophical aspects of the problem. The approach, although little known and usually misunderstood among the high-energy physics (HEP) community, has become the standard way of reasoning in several fields of research and has recently been adopted by the international metrology organizations in their recommendations for assessing measurement uncertainty.

These notes describe a general model for treating uncertainties originating from random and systematic errors in a consistent way and include examples of applications of the model in HEP, e.g. confidence intervals in different contexts, upper/lower limits, treatment of systematic errors, hypothesis tests and unfolding.

3.2 Introduction to the primer

The purpose of a measurement is to determine the value of a physical quantity. One often speaks of the true value, an idealized concept achieved by an infinitely precise and accurate measurement, i.e. immune from errors. In practice the result of a measurement is expressed in terms of the best estimate of the true value and of a related uncertainty. Traditionally the various

[1]These notes are based on lectures given to graduate students in Rome (May 1995) and summer students at DESY (September 1995). The original report is Ref. [27]. In the present report, notes (indicated by **Note added**) are used either for clarification or to refer to those parts not contained in the original primer.

contributions to the overall uncertainty are classified in terms of 'statistical' and 'systematic' uncertainties: expressions which reflect the sources of the experimental errors (the quotation marks indicate that a different way of classifying uncertainties will be adopted here).

Statistical uncertainties arise from variations in the results of repeated observations under (apparently) identical conditions. They vanish if the number of observations becomes very large (*"the uncertainty is dominated by systematics"* is the typical expression used in this case) and can be treated — in most cases, but with some exceptions of great relevance in HEP — using conventional statistics based on the frequency-based definition of probability.

On the other hand, it is not possible to treat systematic uncertainties coherently in the frequentistic framework. Several ad hoc prescriptions for how to combine statistical and systematic uncertainties can be found in textbooks and in the literature: *"add them linearly"*; *"add them linearly if ..., else add them quadratically"*; *"don't add them at all"*, and so on (see, e.g., Part 3 of Ref. [1]). The fashion at the moment is to add them quadratically if they are considered independent, or to build a covariance matrix of statistical and systematic uncertainties to treat general cases. These procedures are not justified by conventional statistical theory, but they are accepted because of the pragmatic good sense of physicists. For example, an experimentalist may be reluctant to add twenty or more contributions linearly to evaluate the uncertainty of a complicated measurement, or decide to treat the correlated systematic uncertainties statistically, in both cases unaware of, or simply not caring about, violating frequentistic principles.

The only way to deal with these and related problems in a consistent way is to abandon the frequentistic interpretation of probability introduced at the beginning of this century, and to recover the intuitive concept of probability as degree of belief. Stated differently, one needs to associate the idea of probability with the lack of knowledge, rather than to the outcome of repeated experiments. This has been recognized also by the International Organization for Standardization (ISO), which assumes the subjective definition of probability in its *"Guide to the expression of uncertainty in measurement"* [3].

This primer is organized as follows:

- Sections 3.3–3.6 give a general introduction to subjective probability.
- Sections 4.1–4.2 summarize some concepts and formulae concerning random variables, needed for many applications.
- Section 5.1 introduces the problem of measurement uncertainty and deals with the terminology.
- Sections 5.2–5.3 present the inferential model.
- Sections 5.4–5.6 show several physical applications of the model.
- Section 6.1 deals with the approximate methods needed when the general solution becomes complicated; in this context the ISO recommendations will be presented and discussed.
- Section 6.2 deals with uncertainty propagation. It is particularly short because, in this scheme, there is no difference between the treatment of systematic uncertainties and indirect measurements; the section simply refers to the results of Sections 5.4–6.1.
- Section 6.3 is dedicated to a detailed discussion about the covariance matrix of correlated data and the trouble it may cause.
- Section 7.1 was added as an example of a more complicated inference (multidimensional unfolding) than those treated in Sections 5.4–6.2.

3.3 Probability

3.3.1 What is probability?

The standard answers to this question are

1. the ratio of the number of favourable cases to the number of all cases;
2. the ratio of the number of times the event occurs in a test series to the total number of trials in the series.

It is very easy to show that neither of these statements can define the concept of probability:

- Definition 1 lacks the clause 'if all the cases are equally probable'. This has been done here intentionally, because people often forget it. The fact that the definition of probability makes use of the term 'probability' is clearly embarrassing. Often in textbooks the clause is replaced by 'if all the cases are equally possible', ignoring that in this context 'possible' is just a synonym of 'probable'. There is no way out. This statement does not define probability but gives, at most, a useful rule for evaluating it – assuming we know what probability is, i.e. of what we are talking about. The fact that this definition is labelled 'classical' or 'Laplace' simply shows that some authors are not aware of what the 'classicals' (Bayes, Gauss, Laplace, Bernoulli, etc.) thought about this matter. We shall call this definition 'combinatorial'.
- Definition 2 is also incomplete, since it lacks the condition that the number of trials must be very large (it goes to infinity). But this is a minor point. The crucial point is that the statement merely defines the relative frequency with which an event (a phenomenon) occurred in the past. To use frequency as a measurement of probability we have to assume that the phenomenon occurred in the past, and will occur in the future, with the same probability. But who can tell if this hypothesis is correct? Nobody: we have to guess in every single case. Note that, while in the first definition the assumption of equal probability was explicitly stated, the analogous clause is often missing from the second one. We shall call this definition 'frequentistic'.

We have to conclude that if we want to make use of these statements to assign a numerical value to probability, in those cases in which we judge that the clauses are satisfied, we need a better definition of probability.

3.3.2 Subjective definition of probability

So, what is probability? Consulting a good dictionary helps. Webster's states, for example, that *"probability is the quality, state, or degree of being probable"*, and then that 'probable' means *"supported by evidence strong enough to make it likely though not certain to be true"*. The concept of probable arises in reasoning when the concept of certain is not applicable. If we cannot state firmly whether an event (we use this word as a synonym for any possible statement, or proposition, relative to past, present or future) is true or false, we just say that it is possible or probable. Different events may have different levels of probability, depending whether we think that they are more likely to be true or false (see Fig. 3.1).

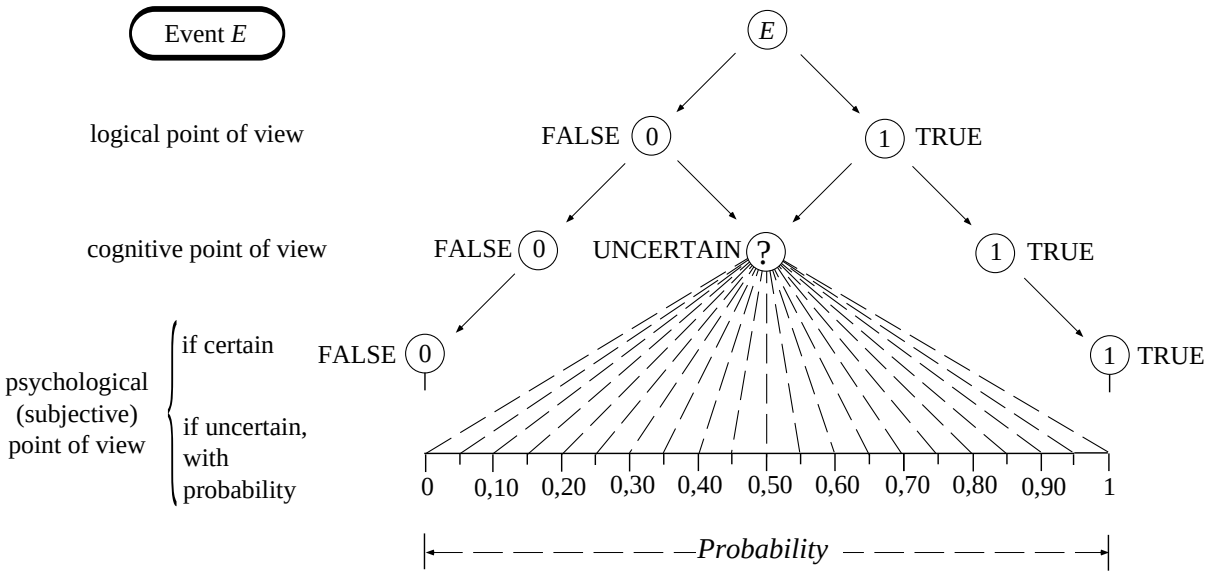

Figure 3.1: Certain and uncertain events [28].

The concept of probability is then simply

a measure of the degree of belief that an event will[2] *occur.*

This is the kind of definition that one finds in Bayesian books (see e.g. Refs. [11, 19, 29, 30, 31]) and the formulation cited here is that given in the ISO Guide [3], which we will discuss later.

At first sight this definition does not seem to be superior to the combinatorial or the frequentistic ones. At least they give some practical rules to calculate 'something'. Defining probability as degree of belief seems too vague to be of any use. We need, then, some explanation of its meaning and a tool to evaluate it with greater precision than intuitive degrees of beliefs can provide. We will look at this tool (Bayes' theorem) later. We will end this section with some explanatory remarks on the definition, but first let us discuss the advantages of this definition.

- It is natural, very general and can be applied to any imaginable event, independently of the feasibility of making an inventory of all (equally) possible and favourable cases, or of repeating the experiment under conditions of equal probability.

- It avoids the linguistic schizophrenia of having to distinguish 'scientific' probability (i.e. strictly based on 'definitions' 1 and 2 of the previous section) from 'non-scientific' probability used in everyday reasoning, including research activity (a meteorologist might feel offended to hear that evaluating the probability of rain tomorrow is not scientific).

- As far as measurements are concerned, it allows us to talk about the probability of the true value of a physical quantity, or of any scientific hypothesis. In the frequentistic frame it is only possible to talk about the probability of the outcome of an experiment, as the true value is considered to be a constant. This approach is so unnatural that most physicists speak of '95% probability that the mass of the top quark is between ...', although they believe that the correct definition of probability is the limit of the frequency.

[2]The use of the future tense does not imply that this definition can only be applied for future events. It simply means that the statement will be proven to be true, even if it refers to the past. Think for example of the probability that it was raining in Rome on the day of the battle of Waterloo.

- It is possible to make a very general theory of uncertainty which can take into account any source of statistical or systematic error, independently of their distribution.

To get a better understanding of the subjective definition of probability let us take a look at odds in betting. The higher the degree of belief that an event will occur, the higher the amount of money A that someone (a rational better) is ready to pay in order to receive a sum of money B if the event occurs. Clearly the bet must be acceptable in both directions ('coherent' is the correct adjective), i.e. the amount of money A must be smaller or equal to B and not negative (who would accept such a bet?). The cases of $A = 0$ and $A = B$ mean that the events are considered to be false or true, respectively, and obviously it is not worth betting on certainty. They are just limit cases, and in fact they can be treated with standard logic. It seems reasonable[3] that the amount of money A that one is willing to pay grows linearly with the degree of belief. It follows that if someone thinks that the probability of the event E is p, then he will bet $A = p\,B$ to get B if the event occurs, and to lose $p\,B$ if it does not. It is easy to demonstrate that the condition of coherence implies that $0 \leq p \leq 1$.

What has gambling to do with physics? The definition of probability through betting odds has to be considered operational, although there is no need to make a bet (with whom?) each time one presents a result. It has the important role of forcing one to make an honest assessment of the value of probability that one believes. One could replace money with other forms of gratification or penalization, like the increase or the loss of scientific reputation. Moreover, the fact that this operational procedure is not to be taken literally should not be surprising. Many physical quantities are defined in a similar way. Think, for example, of the textbook definition of the electric field, and try to use it to measure $\vec{E}$ in the proximity of an electron. A nice example comes from the definition of a poisonous chemical compound: *"it would be lethal if ingested"* [4]. Clearly it is preferable to keep this operational definition at a hypothetical level, even though it is the best definition of the concept.

3.3.3 Rules of probability

The subjective definition of probability, together with the condition of coherence, requires that $0 \leq p \leq 1$. This is one of the rules which probability has to obey. It is possible, in fact, to demonstrate that coherence yields to the standard rules of probability, generally known as axioms. At this point it is worth clarifying the relationship between the axiomatic approach and the others.

- Combinatorial and frequentistic definitions give useful rules for evaluating probability, although they do not, as it is often claimed, define the concept.

- In the axiomatic approach one refrains from defining what the probability is and how to evaluate it: probability is just any real number which satisfies the axioms. It is easy to demonstrate that the probabilities evaluated using the combinatorial and the frequentistic prescriptions do in fact satisfy the axioms.

- The subjective approach to probability, together with the coherence requirement, defines what probability is and provides the rules which its evaluation must obey; these rules turn out to be the same as the axioms.

[3] This is not always true in real life as the importance of a given amount of money differs from person to person. The problem can be solved if the bet is considered *virtual*, i.e. the bet one would consider fair if one had an infinite budget.

[4] Both examples are from R. Scozzafava [32].

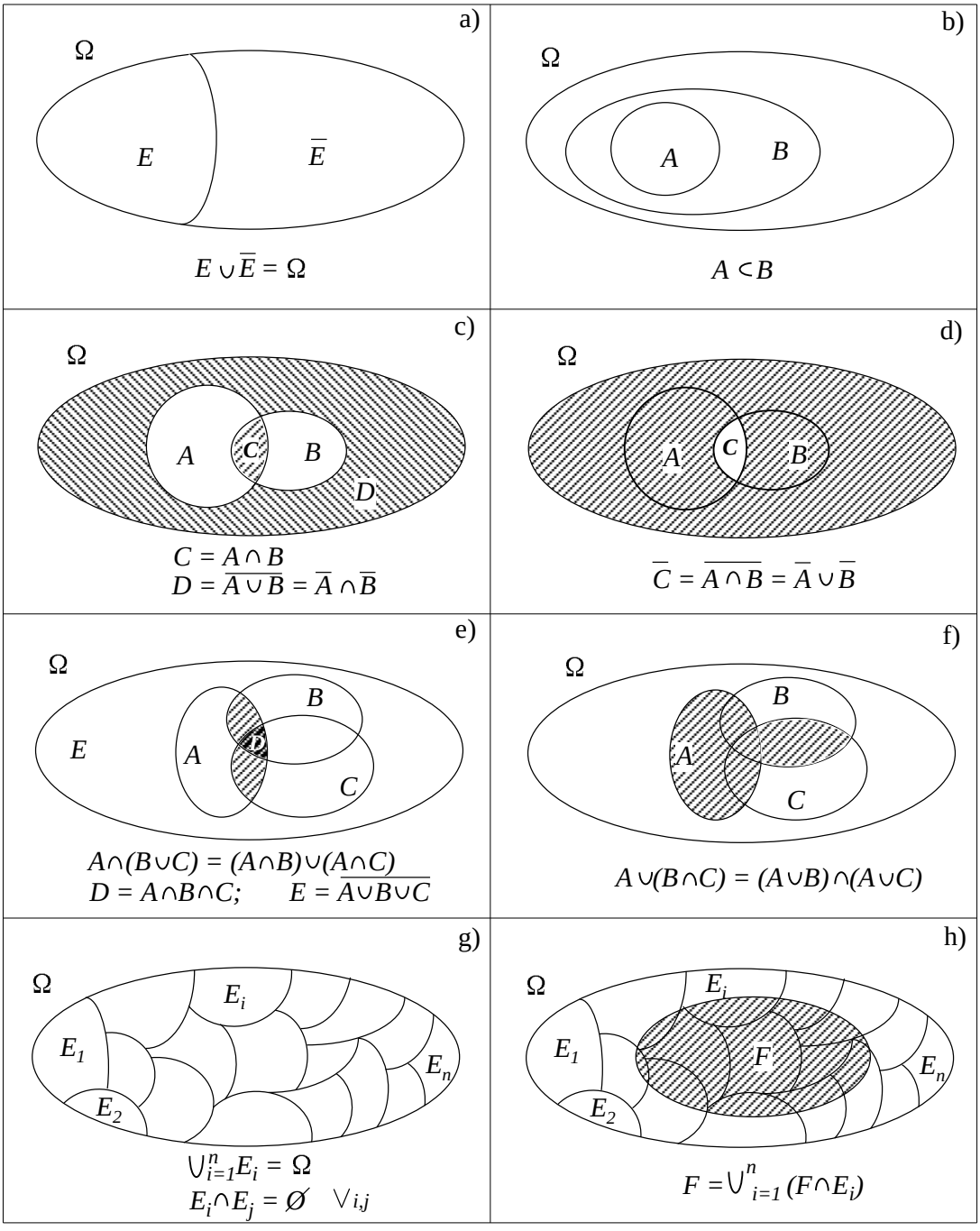

Figure 3.2: Venn diagrams and set properties.

Table 3.1: Events versus sets.

Events	Sets	
		Symbol
event	set	E
certain event	sample space	Ω
impossible event	empty set	$\emptyset$
implication	inclusion (subset)	$E_1 \subseteq E_2$
opposite event (complementary)	complementary set	$\overline{E} \quad (E \cup \overline{E} = \Omega)$
logical product ("AND")	intersection	$E_1 \cap E_2$
logical sum ("OR")	union	$E_1 \cup E_2$
incompatible events	disjoint sets	$E_1 \cap E_2 = \emptyset$
complete class	finite partition	$\begin{cases} E_i \cap E_j = \emptyset \quad \forall\, i \neq j \\ \cup_i E_i = \Omega \end{cases}$

Since everybody is familiar with the axioms and with the analogy *events* $\Leftrightarrow$ *sets* (see Fig. 3.2 and Table 3.3.3) let us remind ourselves of the rules of probability in this form:

Axiom 1 $0 \leq P(E) \leq 1$;

Axiom 2 $P(\Omega) = 1$ (a certain event has probability 1);

Axiom 3 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$, if $E_1 \cap E_2 = \emptyset$.

From the basic rules the following properties can be derived:

1: $P(E) = 1 - P(\overline{E})$;

2: $P(\emptyset) = 0$;

3: if $A \subseteq B$ then $P(A) \leq P(B)$;

4: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

We also anticipate here another rule which will be discussed in Section 3.4.1:

5: $P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(A) \cdot P(B \,|\, A)$.

3.3.4 Subjective probability and objective description of the physical world

The subjective definition of probability seems to contradict the aim of physicists to describe the laws of physics in the most objective way (whatever this means ...). This is one of the reasons why many regard the subjective definition of probability with suspicion (but probably the main reason is because we have been taught at university that probability is frequency). The main philosophical difference between this concept of probability and an objective definition that we would have liked (but which does not exist in reality) is that $P(E)$ is not an intrinsic characteristic of the event E, but depends on the state of information available to whoever evaluates $P(E)$. The ideal concept of objective probability is recovered when everybody has the

Table 3.2: Results of measurements of the gravitational constant G_N.

Institute	$G_N \left(10^{-11} \frac{\mathrm{m}^3}{\mathrm{kg \cdot s}^2}\right)$	$\frac{\sigma(G_N)}{G_N}$ (ppm)	$\frac{G_N - G_N^C}{G_N^C}$ (10^{-3})
CODATA 1986 ("G_N^C")	6.6726 ± 0.0009	128	–
PTB (Germany) 1994	6.7154 ± 0.0006	83	$+6.41 \pm 0.16$
MSL (New Zealand) 1994	6.6656 ± 0.0006	95	-1.05 ± 0.16
Uni-Wuppertal (Germany) 1995	6.6685 ± 0.0007	105	-0.61 ± 0.17

same state of information. But even in this case it would be better to speak of intersubjective probability. The best way to convince ourselves about this aspect of probability is to try to ask practical questions and to evaluate the probability in specific cases, instead of seeking refuge in abstract questions. I find, in fact, that — to paraphrase a famous statement about Time — probability is objective as long as I am not asked to evaluate it. Here are some examples.

Example 1: What is the probability that a molecule of nitrogen at room temperature has a velocity between 400 and 500 m/s? The answer appears easy: Take the Maxwell distribution formula from a textbook, calculate an integral and get a number. Now let us change the question: I give you a vessel containing nitrogen and a detector capable of measuring the speed of a single molecule and you set up the apparatus (or you let a person you trust do it). Now, what is the probability that the first molecule that hits the detector has a velocity between 400 and 500 m/s? Anybody who has minimal experience (direct or indirect) of experiments would hesitate before answering. He would study the problem carefully and perform preliminary measurements and checks. Finally he would probably give not just a single number, but a range of possible numbers compatible with the formulation of the problem. Then he starts the experiment and eventually, after 10 measurements, he may form a different opinion about the outcome of the eleventh measurement.

Example 2: What is the probability that the gravitational constant G_N has a value between $6.6709 \cdot 10^{-11}$ and $6.6743 \cdot 10^{-11}$ $\mathrm{m^3 kg^{-1} s^{-2}}$? Before 1994 you could have looked at the latest issue of the Particle Data Book [33] and answered that the probability was 95%. Since then — as you probably know — three new measurements of G_N have been performed [34] and we now have four numbers which do not agree with each other (see Table 3.3.4). The probability of the true value of G_N being in that range is currently dramatically decreased.

Example 3: What is the probability that the mass of the top quark, or that of any of the supersymmetric particles, is below 20 or 50 GeV/c^2? Currently it looks as if it must be zero. Ten years ago many experiments were intensively looking for these particles in those energy ranges. Because so many people where searching for them, with enormous human and capital investment, it meant that, at that time, the probability was considered rather high: high enough for fake signals to be reported as strong evidence for them.[5]

[5] We will talk later about the influence of *a priori* beliefs on the outcome of an experimental investigation.

The above examples show how the evaluation of probability is conditioned by some *a priori* (theoretical) prejudices and by some facts (experimental data). 'Absolute' probability makes no sense. Even the classical example of probability 1/2 for each of the results in tossing a coin is only acceptable if the coin is regular, it does not remain vertical (not impossible when playing on the beach), it does not fall into a manhole, etc.

The subjective point of view is expressed in a provocative way by de Finetti's [11]

"PROBABILITY DOES NOT EXIST".

3.4 Conditional probability and Bayes' theorem

3.4.1 Dependence of the probability on the state of information

If the state of information changes, the evaluation of the probability also has to be modified. For example most people would agree that the probability of a car being stolen depends on the model, age and parking site. To take an example from physics, the probability that in a HERA detector a charged particle of 1 GeV gives a certain number of ADC counts due to the energy loss in a gas detector can be evaluated in a very general way — using HEP jargon — by making a (huge) Monte Carlo simulation which takes into account all possible reactions (weighted with their cross-sections), all possible backgrounds, changing all physical and detector parameters within reasonable ranges, and also taking into account the trigger efficiency. The probability changes if one knows that the particle is a K^+: instead of very complicated Monte Carlo simulation one can just run a single particle generator. But then it changes further if one also knows the exact gas mixture, pressure, etc., up to the latest determination of the pedestal and the temperature of the ADC module.

3.4.2 Conditional probability

Although everybody knows the formula of conditional probability, it is useful to derive it here.[6] The notation is $P(E\,|\,H)$, to be read 'probability of E given H', where H stands for hypothesis. This means: the probability that E will occur under the hypothesis that H has occurred.[7]

The event $E\,|\,H$ can have three values:

TRUE: if E is TRUE and H is TRUE;

FALSE: if E is FALSE and H is TRUE;

UNDETERMINED: if H is FALSE; in this case we are merely uninterested in what happens to E. In terms of betting, the bet is invalidated and none loses or gains.

Then $P(E)$ can be written $P(E\,|\,\Omega)$, to state explicitly that it is the probability of E whatever happens to the rest of the world (Ω means all possible events). We realize immediately that this condition is really too vague and nobody would bet a penny on a such a statement. The reason for usually writing $P(E)$ is that many conditions are implicitly, and reasonably, assumed

[6] **Note added**: for a further discussion about the meaning of 'the formula of conditional probability' see Section 8.3.

[7] $P(E\,|\,H)$ should not be confused with $P(E \cap H)$, 'the probability that both events occur'. For example $P(E \cap H)$ can be very small, but nevertheless $P(E\,|\,H)$ very high. Think of the limit case

$$P(H) \equiv P(H \cap H) \leq P(H\,|\,H) = 1 \ :$$

'H given H' is a certain event no matter how small $P(H)$ is, even if $P(H) = 0$ (in the sense of Section 4.1.2).

in most circumstances. In the classical problems of coins and dice, for example, one assumes that they are regular. In the example of the energy loss, it was implicit (obvious) that the high voltage was on (at which voltage?) and that HERA was running (under which condition?). But one has to take care: many riddles are based on the fact that one tries to find a solution which is valid under more strict conditions than those explicitly stated in the question[35], and many people make bad business deals by signing contracts in which what was obvious was not explicitly stated.

In order to derive the formula of conditional probability let us assume for a moment that it is reasonable to talk about absolute probability $P(E) = P(E \,|\, \Omega)$, and let us rewrite

$$\begin{aligned} P(E) \equiv P(E \,|\, \Omega) &\underset{\mathbf{a}}{=} P(E \cap \Omega) \\ &\underset{\mathbf{b}}{=} P\left(E \cap (H \cup \overline{H})\right) \\ &\underset{\mathbf{c}}{=} P\left((E \cap H) \cup (E \cap \overline{H})\right) \\ &\underset{\mathbf{d}}{=} P(E \cap H) + P(E \cap \overline{H})\,, \end{aligned} \tag{3.1}$$

where the result has been achieved through the following steps:

(a) E implies Ω (i.e. $E \subseteq \Omega$) and hence $E \cap \Omega = E$;

(b) the complementary events H and $\overline{H}$ make a finite partition of Ω, i.e. $H \cup \overline{H} = \Omega$;

(c) distributive property;

(d) axiom 3.

The final result of (3.1) is very simple: $P(E)$ is equal to the probability that E occurs and H also occurs, plus the probability that E occurs but H does not occur. To obtain $P(E \,|\, H)$ we just get rid of the subset of E which does not contain H (i.e. $E \cap \overline{H}$) and renormalize the probability dividing by $P(H)$, assumed to be different from zero. This guarantees that if $E = H$ then $P(H \,|\, H) = 1$. We get, finally, the well-known formula

$$P(E \,|\, H) = \frac{P(E \cap H)}{P(H)} \qquad [P(H) \neq 0]\,. \tag{3.2}$$

In the most general (and realistic) case, where both E and H are conditioned by the occurrence of a third event $H_\circ$, the formula becomes

$$P(E \,|\, H, H_\circ) = \frac{P\left(E \cap (H \,|\, H_\circ)\right)}{P(H \,|\, H_\circ)} \qquad [P(H \,|\, H_\circ) \neq 0]\,. \tag{3.3}$$

Usually we shall make use of (3.2) (which means $H_\circ = \Omega$) assuming that Ω has been properly chosen. We should also remember that (3.2) can be resolved with respect to $P(E \cap H)$, obtaining the well-known

$$P(E \cap H) = P(E \,|\, H) P(H)\,, \tag{3.4}$$

and by symmetry

$$P(E \cap H) = P(H \,|\, E) P(E)\,. \tag{3.5}$$

We remind that two events are called independent if

$$P(E \cap H) = P(E)P(H)\,. \tag{3.6}$$

This is equivalent to saying that $P(E\,|\,H) = P(E)$ and $P(H\,|\,E) = P(H)$, i.e. the knowledge that one event has occurred does not change the probability of the other. If $P(E\,|\,H) \neq P(E)$ then the events E and H are correlated. In particular:

- if $P(E\,|\,H) > P(E)$ then E and H are positively correlated;
- if $P(E\,|\,H) < P(E)$ then E and H are negatively correlated.

3.4.3 Bayes' theorem

Let us think of all the possible, mutually exclusive, hypotheses H_i which could condition the event E. The problem here is the inverse of the previous one: what is the probability of H_i under the hypothesis that E has occurred? For example, what is the probability that a charged particle which went in a certain direction and has lost between 100 and 120 keV in the detector is a μ, π, K, or p? Our event E is 'energy loss between 100 and 120 keV', and H_i are the four 'particle hypotheses'. This example sketches the basic problem for any kind of measurement: having observed an effect, to assess the probability of each of the causes which could have produced it. This intellectual process is called inference, and it will be discussed in Section 5.2.

In order to calculate $P(H_i\,|\,E)$ let us rewrite the joint probability $P(H_i \cap E)$, making use of (3.4–3.5), in two different ways:

$$P(H_i\,|\,E)P(E) = P(E\,|\,H_i)P(H_i)\,, \tag{3.7}$$

obtaining

$$\boxed{P(H_i\,|\,E) = \frac{P(E\,|\,H_i)P(H_i)}{P(E)}\,,} \tag{3.8}$$

or

$$\boxed{\frac{P(H_i\,|\,E)}{P(H_i)} = \frac{P(E\,|\,H_i)}{P(E)}\,.} \tag{3.9}$$

Since the hypotheses H_i are mutually exclusive (i.e. $H_i \cap H_j = \emptyset$, $\forall\, i,j$) and exhaustive (i.e. $\bigcup_i H_i = \Omega$), E can be written as $\bigcup_i E \cap H_i$, the union of the intersections of E with each of the hypotheses H_i. It follows that

$$\begin{aligned} P(E)\,[\equiv P(E \cap \Omega)] &= P\left(\bigcup_i (E \cap H_i)\right) \\ &= \sum_i P(E \cap H_i) \\ &= \sum_i P(E\,|\,H_i)P(H_i)\,, \end{aligned} \tag{3.10}$$

where we have made use of (3.4) again in the last step. It is then possible to rewrite (3.8) as

$$\boxed{P(H_i\,|\,E) = \frac{P(E\,|\,H_i)P(H_i)}{\sum_j P(E\,|\,H_j)P(H_j)}\,.} \tag{3.11}$$

This is the standard form by which Bayes' theorem is known. (3.8) and (3.9) are also different ways of writing it. As the denominator of (3.11) is nothing but a normalization factor, such that $\sum_i P(H_i \,|\, E) \;=\; 1$, the formula (3.11) can be written as

$$\boxed{P(H_i \,|\, E) \propto P(E \,|\, H_i) P(H_i)\,.} \tag{3.12}$$

Factorizing $P(H_i)$ in (3.11), and explicitly writing that all the events were already conditioned by $H_\circ$, we can rewrite the formula as

$$\boxed{P(H_i \,|\, E, H_\circ) = \alpha P(H_i \,|\, H_\circ)\,,} \tag{3.13}$$

with

$$\alpha = \frac{P(E \,|\, H_i, H_\circ)}{\sum_i P(E \,|\, H_i, H_\circ) P(H_i \,|\, H_\circ)}\,. \tag{3.14}$$

These five ways of rewriting the same formula simply reflect the importance that we shall give to this simple theorem. They stress different aspects of the same concept.

- (3.11) is the standard way of writing it, although some prefer (3.8).
- (3.9) indicates that $P(H_i)$ is altered by the condition E with the same ratio with which $P(E)$ is altered by the condition H_i.
- (3.12) is the simplest and the most intuitive way to formulate the theorem: 'The probability of H_i given E is proportional to the initial probability of H_i times the probability of E given H_i.'
- (3.13–3.14) show explicitly how the probability of a certain hypothesis is updated when the state of information changes:

  $\boxed{P(H_i \,|\, H_\circ)}$ [also indicated as $P_\circ(H_i)$] is the initial, or *a priori*, probability (or simply 'prior') of H_i, i.e. the probability of this hypothesis with the state of information available before the knowledge that E has occurred;

  $\boxed{P(H_i \,|\, E, H_\circ)}$ [or simply $P(H_i \,|\, E)$] is the final, or *a posteriori*, probability of H_i after[8] the new information.

  $\boxed{P(E \,|\, H_i, H_\circ)}$ [or simply $P(E \,|\, H_i)$] is called likelihood.

To better understand the terms 'initial', 'final' and 'likelihood', let us formulate the problem in a way closer to the physicist's mentality, referring to causes and effects: the causes could be all the physical sources which may produce a certain observable (the effect). Using our example of the dE/dx measurement again, the causes are all the possible charged particles which can pass through the detector; the effect is the amount of observed ionization; the likelihoods are the probabilities that each of the particles give that amount of ionization. Note that in this example we have fixed all the other sources of influence: physics process, HERA running conditions, gas mixture, high voltage, track direction, etc. This is our $H_\circ$. The problem immediately gets rather complicated (all real cases, apart from tossing coins and dice, are complicated!). The real inference would be of the kind

$$P(H_i \,|\, E, H_\circ) \propto P(E \,|\, H_i, H_\circ) P(H_i \,|\, H_\circ)\,. \tag{3.15}$$

[8] Note that 'before' and 'after' do not really necessarily imply time ordering, but only the consideration or not of the new piece of information.

For each state $H_\circ$ (the set of all the possible values of the influence parameters) one gets a different result for the final probability. So, instead of getting a single number for the final probability we have a distribution of values. This spread will result in a large uncertainty of $P(H_i \,|\, E)$. This is what every physicist knows: if the calibration constants of the detector and the physics process are not under control, the systematic errors are large and the result is of poor quality.[9]

3.4.4 Conventional use of Bayes' theorem

Bayes' theorem follows directly from the rules of probability, and it can be used in any kind of approach. Let us take an example:

Problem 1: A particle detector has a μ identification efficiency of 95%, and a probability of identifying a π as a μ of 2%. If a particle is identified as a μ, then a trigger is fired. Knowing that the particle beam is a mixture of 90% π and 10% μ, what is the probability that a trigger is really fired by a μ? What is the signal-to-noise (S/N) ratio?

Solution: The two hypotheses (causes) which could condition the event (effect) T (= trigger fired) are μ and π. They are incompatible (clearly) and exhaustive (90% + 10% = 100%). Then:

$$\begin{aligned} P(\mu \,|\, T) &= \frac{P(T \,|\, \mu) P_\circ(\mu)}{P(T \,|\, \mu) P_\circ(\mu) + P(T \,|\, \pi) P_\circ(\pi)} \\ &= \\ &= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.02 \times 0.9} = 0.84\,, \end{aligned} \tag{3.16}$$

and $P(\pi \,|\, T) = 0.16$.

The S/N ratio is $P(\mu \,|\, T)/P(\pi \,|\, T) = 5.3$. It is interesting to rewrite the general expression of the S/N ratio if the effect E is observed as

$$S/N = \frac{P(S \,|\, E)}{P(N \,|\, E)} = \frac{P(E \,|\, S)}{P(E \,|\, N)} \cdot \frac{P_\circ(S)}{P_\circ(N)}\,. \tag{3.17}$$

This formula explicitly shows that when there are noisy conditions,

$$P_\circ(S) \ll P_\circ(N),$$

the experiment must be very selective,

$$P(E \,|\, S) \gg P(E \,|\, N),$$

in order to have a decent S/N ratio.
(How does S/N change if the particle has to be identified by two independent detectors in order to give the trigger? Try it yourself, the answer is $S/N = 251$.)

[9] Formally, the influence of the uncertainty about $H_\circ$ on $P(H_i)$ can be seen in the following way. Indicating by $H_{\circ_j}$ all possible configurations of $H_\circ$, we get from the rules of probability:

$$P(H_i \,|\, E) = \sum_j P(H_i \,|\, E, H_{\circ_j}) P(H_{\circ_j})\,.$$

Problem 2: Three boxes contain two rings each, but in one of them they are both gold, in the second both silver, and in the third one of each type. You have the choice of randomly extracting a ring from one of the boxes, the content of which is unknown to you. You look at the extracted ring, and you then have the possibility of extracting a second ring, again from any of the three boxes. Let us assume the first ring you extract is a gold one. Is it then preferable to extract the second one from the same or from a different box?

Solution: Choosing the same box you have a 2/3 probability of getting a second gold ring. (Try to apply the theorem, or help yourself with intuition; the solution is given in Section 8.10.)

The difference between the two problems, from the conventional statistics point of view, seems to be the following. In the frequentistic approach only the first problem is meaningful, since the probabilities entering in the problem are evaluated from experimental frequencies. In a pure combinatorial approach only the second problem has a solution. Nevertheless, the question is a little more subtle. What is, for example, the meaning of the 84% probability obtained as the solution of the first problem? It is no longer a ratio between the number of occurrencies of the event and the number of experimental trials. Therefore, strictly speaking, it is not a probability according to the frequentistic 'definition'. It is easy to understand that the only consistent way to interpret such a result is to consider it as the degree of belief that the particle was a muon. The same is true for the solution of the second problem.

In conclusion, although the rules of probability are the same in the different approaches (and therefore also in Bayes' theorem), only in the subjective approach are the results of the calculations consistent at every step with the definition of probability.

3.4.5 Bayesian statistics: learning by experience

The advantage of the Bayesian approach (leaving aside the 'little philosophical detail' of trying to define what probability is) is that one may talk about the probability of any kind of event, as already emphasized. Moreover, the procedure of updating the probability with increasing information is very similar to that followed by the mental processes of rational people.[10] Let us consider a few examples of 'Bayesian use' of Bayes' theorem.

Example 1: Imagine some persons listening to a common friend having a phone conversation with an unknown person X_i, and who are trying to guess who X_i is. Depending on the knowledge they have about the friend, on the language spoken, on the tone of voice, on the subject of conversation, etc., they will attribute some probability to several possible persons. As the conversation goes on they begin to consider some possible candidates for X_i, discarding others, then hesitating perhaps only between a couple of possibilities, until the state of information I is such that they are practically sure of the identity of X_i. This experience has happened to most of us, and it is not difficult to recognize the Bayesian scheme:

$$P(X_i \,|\, I, I_\circ) \propto P(I \,|\, X_i, I_\circ) P(X_i \,|\, I_\circ)\,. \tag{3.18}$$

We have put the initial state of information $I_\circ$ explicitly in (3.18) to remind us that likelihoods and initial probabilities depend on it. If we know nothing about the person, the final probabilities will be very vague, i.e. for many persons X_i the probability will be different from zero, without necessarily favouring any particular person.

[10] **Note added**: Ref. [36] shows an interesting investigations on the relation between perception and Bayesian inference.

Example 2: A person X meets an old friend F in a pub. F proposes that the drinks should be payed for by whichever of the two extracts the card of lower value from a pack (according to some rule which is of no interest to us). X accepts and F wins. This situation happens again in the following days and it is always X who has to pay. What is the probability that F has become a cheat, as the number of consecutive wins n increases?

The two hypotheses are: cheat (C) and honest (H). $P_\circ(C)$ is low because F is an old friend, but certainly not zero: let us assume 5%. To make the problem simpler let us make the approximation that a cheat always wins (not very clever...): $P(W_n \,|\, C) = 1$. The probability of winning if he is honest is, instead, given by the rules of probability assuming that the chance of winning at each trial is 1/2 (why not?, we shall come back to this point later): $P(W_n \,|\, H) = 2^{-n}$. The result

$$P(C \,|\, W_n) = \frac{P(W_n \,|\, C) \cdot P_\circ(C)}{P(W_n \,|\, C) \cdot P_\circ(C) + P(W_n \,|\, H) \cdot P_\circ(H)} \tag{3.19}$$

$$= \frac{1 \cdot P_\circ(C)}{1 \cdot P_\circ(C) + 2^{-n} \cdot P_\circ(H)} \tag{3.20}$$

is shown in the following table.

n	$P(C \,\|\, W_n)$ (%)	$P(H \,\|\, W_n)$ (%)
0	5.0	95.0
1	9.5	90.5
2	17.4	82.6
3	29.4	70.6
4	45.7	54.3
5	62.7	37.3
6	77.1	22.9
...	...	...

Naturally, as F continues to win the suspicion of X increases. It is important to make two remarks.

- The answer is always probabilistic. X can never reach absolute certainty that F is a cheat, unless he catches F cheating, or F confesses to having cheated. This is coherent with the fact that we are dealing with random events and with the fact that any sequence of outcomes has the same probability (although there is only one possibility over 2^n in which F is always luckier). Making use of $P(C \,|\, W_n)$, X can make a decision about the next action to take:
  - continue the game, with probability $P(C \,|\, W_n)$ of losing with certainty the next time too;
  - refuse to play further, with probability $P(H \,|\, W_n)$ of offending the innocent friend.
- If $P_\circ(C) = 0$ the final probability will always remain zero: if X fully trusts F, then he has just to record the occurrence of an *a priori* rare event when n becomes large.

To better follow the process of updating the probability when new experimental data become available, according to the Bayesian scheme

"the final probability of the present inference is the initial probability of the next one."

Let us call $P(C\,|\,W_{n-1})$ the probability assigned after the previous win. The iterative application of the Bayes formula yields

$$P(C\,|\,W_n) = \frac{P(W\,|\,C)\cdot P(C\,|\,W_{n-1})}{P(W\,|\,C)\cdot P(C\,|\,W_{n-1}) + P(W\,|\,H)\cdot P(H\,|\,W_{n-1})} \tag{3.21}$$

$$= \frac{1\cdot P(C\,|\,W_{n-1})}{1\cdot P(C\,|\,W_{n-1}) + \frac{1}{2}\cdot P(H\,|\,W_{n-1})}\,, \tag{3.22}$$

where $P(W\,|\,C)=1$ and $P(W\,|\,H)=1/2$ are the probabilities of each win. The interesting result is that exactly the same values of $P(C\,|\,W_n)$ of (3.20) are obtained (try to believe it!).

It is also instructive to see the dependence of the final probability on the initial probabilities, for a given number of wins n.

$P_\circ(C)$	$P(C\,\|\,W_n)$ (%)			
	$n=5$	$n=10$	$n=15$	$n=20$
1 %	24	91	99.7	99.99
5 %	63	98	99.94	99.998
50 %	97	99.90	99.997	99.9999

As the number of experimental observations increases the conclusions no longer depend, practically, on the initial assumptions. This is a crucial point in the Bayesian scheme and it will be discussed in more detail later.

3.5 Hypothesis test (discrete case)

Although in conventional statistics books this argument is usually dealt with in one of the later chapters, in the Bayesian approach it is so natural that it is in fact the first application, as we have seen in the above examples. We summarize here the procedure:

- probabilities are attributed to the different hypotheses using initial probabilities and experimental data (via the likelihood);
- the person who makes the inference — or the 'user' — will make a decision for which he is fully responsible.

If one needs to compare two hypotheses, as in the example of the S/N calculation, the ratio of the final probabilities can be taken as a quantitative result of the test. Let us rewrite the S/N formula (3.17) in the most general case:

$$\frac{P(H_1\,|\,E,H_\circ)}{P(H_2\,|\,E,H_\circ)} = \frac{P(E\,|\,H_1,H_\circ)}{P(E\,|\,H_2,H_\circ)}\cdot\frac{P(H_1\,|\,H_\circ)}{P(H_2\,|\,H_\circ)}\,, \tag{3.23}$$

where again we have reminded ourselves of the existence of $H_\circ$. The ratio depends on the product of two terms: the ratio of the priors and the ratio of the likelihoods. When there is absolutely

no reason for choosing between the two hypotheses the prior ratio is 1 and the decision depends only on the other term, called 'the Bayes factor'. If one firmly believes in either hypothesis, the Bayes factor is of minor importance, unless it is zero or infinite (i.e. one and only one of the likelihoods is vanishing). Perhaps this is disappointing for those who expected objective certainty from a probability theory, but this is in the nature of things.

3.6 Choice of the initial probabilities (discrete case)

3.6.1 General criteria

The dependence of Bayesian inferences on initial probability is considered by opponents as the fatal flaw in the theory. But this criticism is less severe than one might think at first sight.[11] In fact:

- It is impossible to construct a theory of uncertainty which is not affected by this 'illness'. Those methods which are advertised as being 'objective' tend in reality to hide the hypotheses on which they are grounded. A typical example is the maximum likelihood method, of which we will talk later.
- As the amount of information increases the dependence on initial prejudices diminishes.
- When the amount of information is very limited, or completely lacking, there is nothing to be ashamed of if the inference is dominated by *a priori* assumptions.

It is well known to all experienced physicists that conclusions drawn from an experimental result (and sometimes even the result itself!) often depend on prejudices about the phenomenon under study. Some examples:

- When doing quick checks on a device, a single measurement is usually performed if the value is 'what it should be', but if it is not then many measurements tend to be made.
- Results are sometimes influenced by previous results or by theoretical predictions. See for example Fig. 3.3 taken from the Particle Data Book [33]. The interesting book *"How experiments end"* [37] discusses, among others, the issue of when experimentalists are happy with the result and stop correcting for the systematics.
- Slight deviations from the background might be interpreted as a signal (e.g. as for the first claim of discovery of the top quark in spring 1994), while larger signals are viewed with suspicion if they are unwanted by the physics establishment.[12]
- Experiments are planned and financed according to the prejudices of the moment.[13]

These comments are not intended to justify unscrupulous behaviour or sloppy analysis. They are intended, instead, to remind us — if need be — that scientific research is ruled by subjectivity much more than outsiders imagine. The transition from subjectivity to objectivity begins when there is a large consensus among the most influential people about how to interpret the results.[14]

[11] **Note added**: for an extensive discussion about priors see Ref. [22].

[12] A case, concerning the search for electron compositeness in $e^+ e^-$ collisions, is discussed in Ref. [38].

[13] For a recent delightful report, see Ref. [39].

[14] *"A theory needs to be confirmed by experiments. But it is also true that an experimental result needs to be confirmed by a theory."* This sentence expresses clearly — though paradoxically — the idea that it is difficult to accept a result which is not rationally justified. An example of results not confirmed by the theory are the R measurements in deep-inelastic scattering shown in Fig. 3.4. Given the conflict in this situation, physicists tend to believe more in QCD and use the 'low-x' extrapolations (of what?) to correct the data for the unknown values of R.

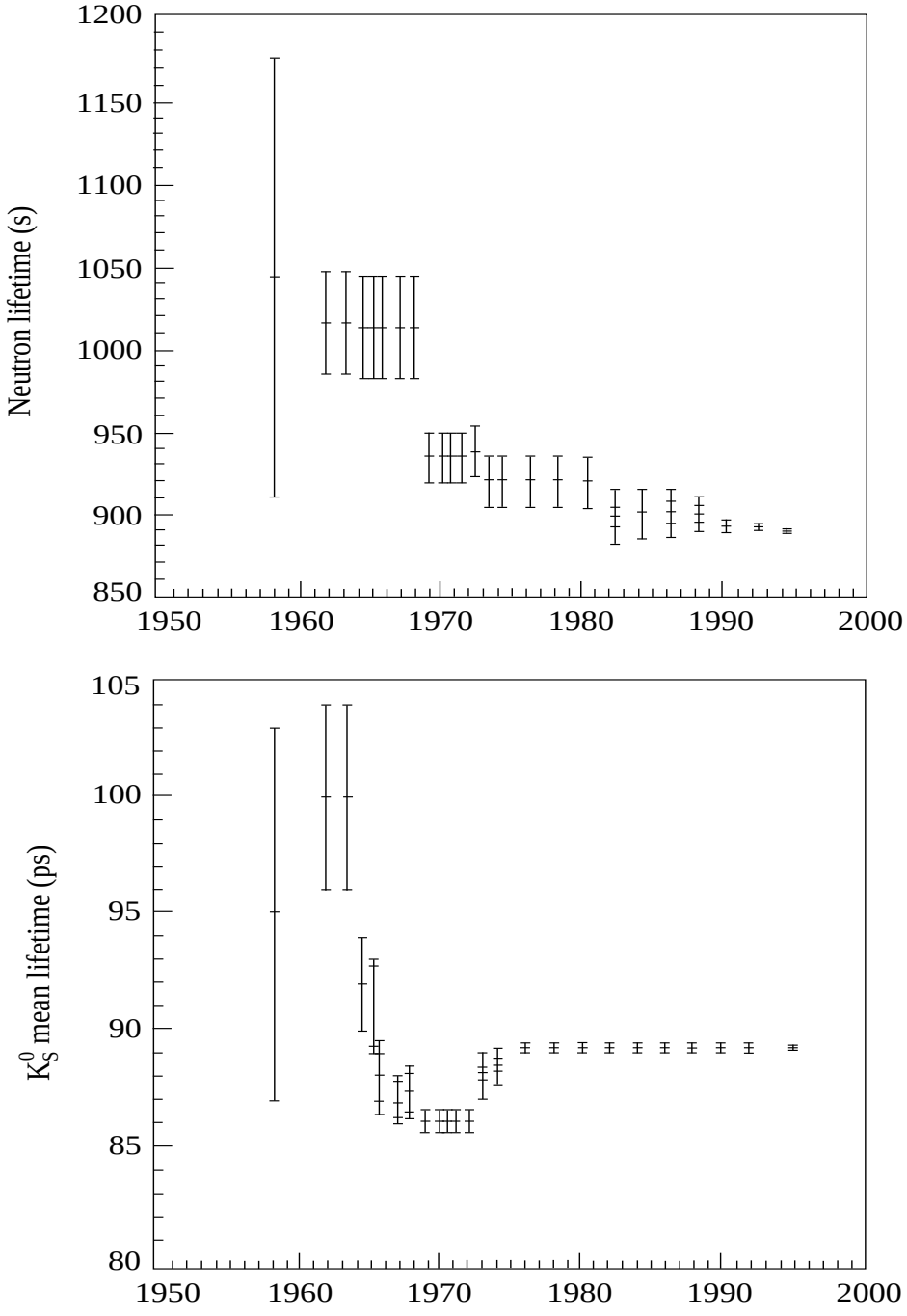

Figure 3.3: Results on two physical quantities as a function of the publication date.

In this context, the subjective approach to statistical inference at least teaches us that every assumption must be stated clearly and all available information which could influence conclusions must be weighed with the maximum attempt at objectivity.[15]

What are the rules for choosing the 'right' initial probabilities? As one can imagine, this is an open and debated question among scientists and philosophers. My personal point of view is that one should avoid pedantic discussion of the matter, because the idea of universally true priors reminds me terribly of the famous 'angels' sex' debates.

If I had to give recommendations, they would be the following.

- The *a priori* probability should be chosen in the same spirit as the rational person who places a bet, seeking to minimize the risk of losing.
- General principles — like those that we will discuss in a while — may help, but since it may be difficult to apply elegant theoretical ideas in all practical situations, in many circumstances the guess of the expert can be relied on for guidance.
- To avoid using as prior the results of other experiments dealing with the same open problem, otherwise correlations between the results would prevent all comparison between the experiments and thus the detection of any systematic errors. I find that this point is generally overlooked by statisticians.

3.6.2 Insufficient reason and maximum entropy

The first and most famous criterion for choosing initial probabilities is the simple 'Principle of Insufficient Reason' (or 'Indifference Principle'): If there is no reason to prefer one hypothesis over alternatives, simply attribute the same probability to all of them. This was stated as a principle by Laplace[16] in contrast to Leibnitz' famous 'Principle of Sufficient Reason', which, in simple words, states that 'nothing happens without a reason'. The indifference principle applied to coin and die tossing, to card games or to other simple and symmetric problems, yields to the well-known rule of probability evaluation that we have called combinatorial. Since it is impossible not to agree with this point of view, in the cases for which one judges that it does apply, the combinatorial definition of probability is recovered in the Bayesian approach if the word 'definition' is simply replaced by 'evaluation rule'. We have in fact already used this reasoning in previous examples.

A modern and more sophisticated version of the Indifference Principle is the Maximum Entropy Principle. The information entropy function of n mutually exclusive events, to each of which a probability p_i is assigned, is defined as [40]

$$H(p_1, p_2, \ldots p_n) = -K \sum_{i=1}^{n} p_i \ln p_i \,, \tag{3.24}$$

with K a positive constant. The principle states that *"in making inferences on the basis of partial information we must use that probability distribution which has the maximum entropy subject to whatever is known"* [41]. Note that, in this case, 'entropy' is synonymous with 'uncertainty' [41]. One can show that, in the case of absolute ignorance about the events E_i, the maximization of

[15] It may look paradoxical, but, due to the normative role of the coherent bet, subjective assessments are more objective than using, without direct responsibility, someone else's formulae. For example, even the knowledge that somebody else has a different evaluation of the probability is new information which must be taken into account.

[16] It may help in understanding Laplace's approach if we consider that he called the theory of probability *"good sense turned into calculation."*

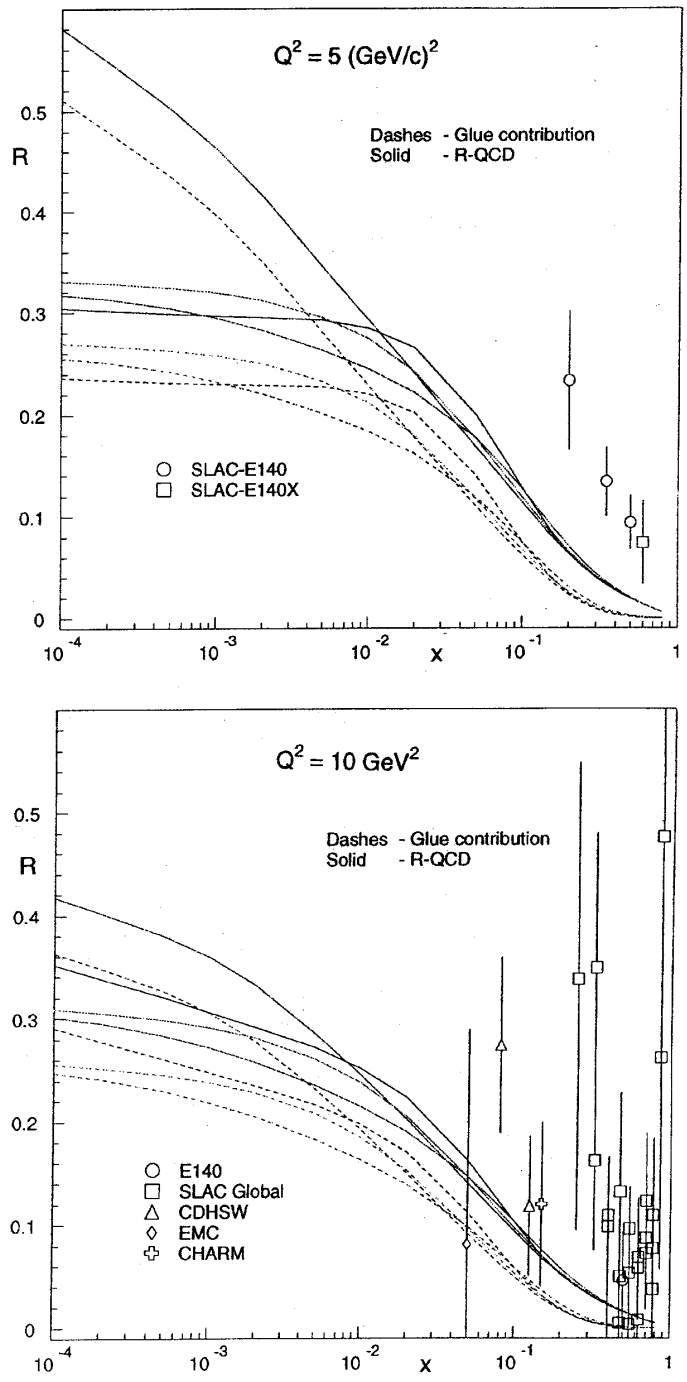

Figure 3.4: $R = \sigma_L/\sigma_T$ as a function of the deep-inelastic scattering variable x as measured by experiments and as predicted by QCD.

the information uncertainty, with the constraint that $\sum_{i=1}^{n} p_i = 1$, yields the classical $p_i = 1/n$ (any other result would have been worrying ...).

Although this principle is sometimes used in combination with the Bayes formula for inferences (also applied to measurement uncertainty, see Ref. [23]), it will not be used for applications in these notes. Those who are interested in entropy, both in information and in probability theory can find a clear introduction in Ref. [42].

Chapter 4

Distributions (a concise reminder)

4.1 Random variables

In the discussion which follows I will assume that the reader is familiar with random variables, distributions, probability density functions, and expected values, as well as with the most frequently used distributions. This section is only intended as a summary of concepts and as a presentation of the notation used in the subsequent sections.

4.1.1 Discrete variables

Uncertain numbers are numbers in respect of which we are in a condition of uncertainty. They can be the number associated with the outcome of a die, to the number which will be read on a scale when a measurement is performed, or to the numerical value of a physics quantity. In the following, we will call uncertain numbers also random variables, to come close to what physicists are used to, but one should not think, then, that random variables are only associated with the outcomes of repeated experiments. Stated simply, to define a random variable X means to find a rule which allows a real number to be related univocally (but not necessarily biunivocal) to an event (E). One could write this expression $X(E)$. Discrete variables assume a countable range, finite or not. We shall indicate the variable with X and its numerical realization with x; and differently from other notations, the symbol x (in place of n or k) is also used for discrete variables.

Here is a list of definitions, properties and notations.

Probability function.

To each possible value of X we associate a degree of belief:

$$f(x) = P(X = x)\,. \tag{4.1}$$

$f(x)$, being a probability, must satisfy the following properties:

$$0 \leq f(x_i) \leq 1\,, \tag{4.2}$$

$$P(X = x_i \cup X = x_j) = f(x_i) + f(x_j)\,, \tag{4.3}$$

$$\sum_i f(x_i) = 1\,. \tag{4.4}$$

Cumulative distribution function.

$$F(x_k) \equiv P(X \leq x_k) = \sum_{x_i \leq x_k} f(x_i)\,. \tag{4.5}$$

Properties:

$$F(-\infty) = 0, \tag{4.6}$$
$$F(+\infty) = 1, \tag{4.7}$$
$$F(x_i) - F(x_{i-1}) = f(x_i), \tag{4.8}$$
$$\lim_{\epsilon \to o} F(x+\epsilon) = F(x) \qquad (\textit{right side continuity})\,. \tag{4.9}$$

Expected value (mean).

$$\mu \equiv \mathrm{E}[X] = \sum_i x_i f(x_i)\,. \tag{4.10}$$

In general, given a function $g(X)$ of X,

$$\mathrm{E}[g(X)] = \sum_i g(x_i) f(x_i)\,. \tag{4.11}$$

$\mathrm{E}[\cdot]$ is a linear operator:

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b\,. \tag{4.12}$$

Variance and standard deviation.

Variance:

$$\sigma^2 \equiv \mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2\,. \tag{4.13}$$

Standard deviation:

$$\sigma = \sqrt{\sigma^2}\,. \tag{4.14}$$

Transformation properties:

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)\,, \tag{4.15}$$
$$\sigma(aX+b) = |a|\,\sigma(X)\,. \tag{4.16}$$

Binomial distribution.

$X \sim \mathcal{B}_{n,p}$ (hereafter '$\sim$' stands for 'follows'); $\mathcal{B}_{n,p}$ indicates a binomial with parameters n (integer) and p (real):

$$f(x \,|\, \mathcal{B}_{n,p}) = \frac{n!}{(n-x)!\,x!}\,p^x\,(1-p)^{n-x}\,, \qquad \begin{cases} n = 1, 2, \ldots, \infty \\ 0 \le p \le 1 \\ x = 0, 1, \ldots, n \end{cases}\,. \tag{4.17}$$

Expected value, standard deviation and variation coefficient:

$$\mu = n\,p, \tag{4.18}$$
$$\sigma = \sqrt{n\,p\,(1-p)}, \tag{4.19}$$
$$v \equiv \frac{\sigma}{\mu} = \frac{\sqrt{n\,p\,(1-p)}}{n\,p} \propto \frac{1}{\sqrt{n}}\,. \tag{4.20}$$

$1-p$ is often indicated by q.

Poisson distribution.

$X \sim \mathcal{P}_\lambda$:

$$f(x \,|\, \mathcal{P}_\lambda) = \frac{\lambda^x}{x!}\, e^{-\lambda} \qquad \begin{cases} 0 < \lambda < \infty \\ x = 0, 1, \ldots, \infty \end{cases} . \tag{4.21}$$

(x is an integer, λ is real.)
Expected value, standard deviation and variation coefficient:

$$\mu = \lambda, \tag{4.22}$$
$$\sigma = \sqrt{\lambda}, \tag{4.23}$$
$$v = \frac{1}{\sqrt{\lambda}}. \tag{4.24}$$

Binomial → Poisson.

$$\mathcal{B}_{n,p} \xrightarrow[\substack{n \to \text{`}\infty\text{'} \\ p \to \text{`}0\text{'} \\ (\lambda = n\,p)}]{} P_\lambda$$

4.1.2 Continuous variables: probability density function

Moving from discrete to continuous variables there are the usual problems with infinite possibilities, similar to those found in Zeno's 'Achilles and the tortoise' paradox. In both cases the answer is given by infinitesimal calculus. But some comments are needed:

- The probability of each of the realizations of X is zero ($P(X = x) = 0$); but this does not mean that each value is impossible, otherwise it would be impossible to get any result.
- Although all values x have zero probability, one usually assigns different degrees of belief to them, quantified by the probability density function $f(x)$. Writing $f(x_1) > f(x_2)$, for example, indicates that our degree of belief in x_1 is greater than that in x_2.
- The probability that a random variable lies inside a finite interval, for example $P(a \leq X \leq b)$, is instead finite. If the distance between a and b becomes infinitesimal, then the probability becomes infinitesimal too. If all the values of X have the same degree of belief (and not only equal numerical probability $P(x) = 0$) the infinitesimal probability is simply proportional to the infinitesimal interval $\mathrm{d}P = k\,\mathrm{d}x$. In the general case the ratio between two infinitesimal probabilities around two different points will be equal to the ratio of the degrees of belief in the points (this argument implies the continuity of $f(x)$ on either side of the values). It follows that $\mathrm{d}P = f(x)\,\mathrm{d}x$ and then

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x\,. \tag{4.25}$$

- $f(x)$ has a dimension inverse to that of the random variable.

After this short introduction, here is a list of definitions, properties and notations:

Cumulative distribution function.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'\,, \tag{4.26}$$

or

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}. \tag{4.27}$$

Properties of $f(x)$ and $F(x)$.

- $f(x) \geq 0$,
- $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$,
- $0 \leq F(x) \leq 1$,
- $P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x = \int_{-\infty}^{b} f(x)\,\mathrm{d}x - \int_{-\infty}^{a} f(x)\,\mathrm{d}x = F(b) - F(a)$,
- if $x_2 > x_1$ then $F(x_2) \geq F(x_1)$,
- $\lim_{x\to-\infty} F(x) = 0$,
  $\lim_{x\to+\infty} F(x) = 1$.

Expected value.

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\,f(x)\,\mathrm{d}x, \tag{4.28}$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)\,f(x)\,\mathrm{d}x. \tag{4.29}$$

Uniform distribution. [1]

$X \sim \mathcal{K}(a,b)$:

$$f(x\,|\,\mathcal{K}(a,b)) = \frac{1}{b-a} \quad (a \leq x \leq b), \tag{4.30}$$

$$F(x\,|\,\mathcal{K}(a,b)) = \frac{x-a}{b-a}\,. \tag{4.31}$$

Expected value and standard deviation:

$$\mu = \frac{a+b}{2}, \tag{4.32}$$

$$\sigma = \frac{b-a}{\sqrt{12}}\,. \tag{4.33}$$

[1] The symbols of the following distributions have the parameters within parentheses to indicate that the variables are continuous.

Normal (Gaussian) distribution.

$X \sim \mathcal{N}(\mu, \sigma)$:

$$f(x \,|\, \mathcal{N}(\mu,\sigma)) = \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \left\{ \begin{array}{l} -\infty < \mu < +\infty \\ 0 < \sigma < \infty \\ -\infty < x < +\infty \end{array} \right. , \tag{4.34}$$

where μ and σ (both real) are the expected value and standard deviation,[2] respectively.

Standard normal distribution.

The particular normal distribution of mean 0 and standard deviation 1, usually indicated by Z:

$$Z \sim \mathcal{N}(0,1)\,. \tag{4.35}$$

Exponential distribution.

$T \sim \mathcal{E}(\tau)$:

$$f(t \,|\, \mathcal{E}(\tau)) = \frac{1}{\tau}\, e^{-t/\tau} \qquad \left\{ \begin{array}{l} 0 \le \tau < \infty \\ 0 \le t < \infty \end{array} \right. \tag{4.36}$$

$$F(t \,|\, \mathcal{E}(\tau)) = 1 - e^{-t/\tau}. \tag{4.37}$$

We use the symbol t instead of x because this distribution will be applied to the time domain.

Survival probability:

$$P(T > t) = 1 - F(t \,|\, \mathcal{E}(\tau)) = e^{-t/\tau}. \tag{4.38}$$

Expected value and standard deviation:

$$\mu = \tau \tag{4.39}$$

$$\sigma = \tau. \tag{4.40}$$

The real parameter τ has the physical meaning of lifetime.

Poisson $\leftrightarrow$ Exponential.

If X (= number of counts during the time Δt) is Poisson distributed then T (= interval of time to wait — starting from any instant — before the first count is recorded) is exponentially distributed:

$$X \sim f(x \,|\, \mathcal{P}_\lambda) \qquad \Longleftrightarrow \qquad T \sim f(x \,|\, \mathcal{E}(\tau)) \tag{4.41}$$

$$(\tau = \tfrac{\Delta T}{\lambda}) \quad . \tag{4.42}$$

[2]Mathematicians and statisticians prefer to take σ^2, instead of σ, as second parameter of the normal distribution. Here the standard deviation is preferred, since it is homogeneous to μ and it has a more immediate physical interpretation. So, one has to pay attention to be sure about the meaning of expressions like $\mathcal{N}(0.5, 0.8)$.

4.1.3 Distribution of several random variables

We only consider the case of two continuous variables (X and Y). The extension to more variables is straightforward. The infinitesimal element of probability is $\mathrm{d}F(x,y) = f(x,y)\,\mathrm{d}x\,\mathrm{d}y$, and the probability density function

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}\,. \tag{4.43}$$

The probability of finding the variable inside a certain area A is

$$\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{4.44}$$

Marginal distributions.

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y, \tag{4.45}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,. \tag{4.46}$$

The subscripts X and Y indicate that $f_X(x)$ and $f_Y(y)$ are functions only of X and Y, respectively (to avoid fooling around with different symbols to indicate the generic function), but in most cases we will drop the subscripts if the context helps in resolving ambiguities.

Conditional distributions.

$$f_X(x\,|\,y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int f(x,y)\,\mathrm{d}x}, \tag{4.47}$$

$$f_Y(y\,|\,x) = \frac{f(x,y)}{f_X(x)}, \tag{4.48}$$

$$f(x,y) = f_X(x\,|\,y)\,f_Y(y) \tag{4.49}$$

$$= f_Y(y\,|\,x)\,f_X(x)\,. \tag{4.50}$$

Independent random variables.

$$f(x,y) = f_X(x) f_Y(y) \tag{4.51}$$

(it implies $f_X(x\,|\,y) = f_X(x)$ and $f_Y(y\,|\,x) = f_Y(y)$.)

Bayes' theorem for continuous random variables.

$$\boxed{f(h\,|\,e) = \frac{f(e\,|\,h)\,f_h(h)}{\int f(e\,|\,h)\,f_h(h)\,\mathrm{d}h}\,.} \tag{4.52}$$

(**Note added**: see proof in Section 2.7.)

Expected value.

$$\mu_X = \mathrm{E}[X] = \iint_{-\infty}^{+\infty} x\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{4.53}$$

$$= \int_{-\infty}^{+\infty} x\,f_X(x)\,\mathrm{d}x\,, \tag{4.54}$$

and analogously for Y. In general

$$\mathrm{E}[g(X,Y)] = \int\!\!\int_{-\infty}^{+\infty} g(x,y)\, f(x,y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{4.55}$$

Variance.

$$\sigma_X^2 = \mathrm{E}[X^2] - \mathrm{E}^2[X]\,, \tag{4.56}$$

and analogously for Y.

Covariance.

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}\left[\,(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])\,\right] &(4.57)\\ &= \mathrm{E}[XY] - \mathrm{E}[X] \cdot \mathrm{E}[Y]\,. &(4.58)\end{aligned}$$

If X and Y are independent, then $\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$ and hence $\mathrm{Cov}(X,Y) = 0$ (the opposite is true only if $X,\, Y \sim \mathcal{N}(\cdot)$).

Correlation coefficient.

$$\begin{aligned} \rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} &(4.59)\\ &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}\,. &(4.60)\end{aligned}$$

$$(-1 \le \rho \le 1)$$

Linear combinations of random variables.

If $Y = \sum_i c_i X_i$, with c_i real, then

$$\begin{aligned} \mu_Y = \mathrm{E}[Y] &= \sum_i c_i\, \mathrm{E}[X_i] = \sum_i c_i\, \mu_i, &(4.61)\\ \sigma_Y^2 = \mathrm{Var}(Y) &= \sum_i c_i^2\, \mathrm{Var}(X_i) + 2 \sum_{i<j} c_i\, c_j\, \mathrm{Cov}(X_i, X_j) &(4.62)\\ &= \sum_i c_i^2\, \mathrm{Var}(X_i) + \sum_{i \neq j} c_i\, c_j\, \mathrm{Cov}(X_i, X_j) &(4.63)\\ &= \sum_i c_i^2\, \sigma_i^2 + \sum_{i \neq j} \rho_{ij}\, c_i\, c_j\, \sigma_i\, \sigma_j &(4.64)\\ &= \sum_{ij} \rho_{ij}\, c_i\, c_j\, \sigma_i\, \sigma_j &(4.65)\\ &= \sum_{ij} c_i\, c_j\, \sigma_{ij}\,. &(4.66)\end{aligned}$$

σ_Y^2 has been written in different ways, with increasing levels of compactness, that can be found in the literature. In particular, (4.65) and (4.66) use the notations $\sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j) = \rho_{ij}\, \sigma_i\, \sigma_j$ and $\sigma_{ii} = \sigma_i^2$, and the fact that, by definition, $\rho_{ii} = 1$.

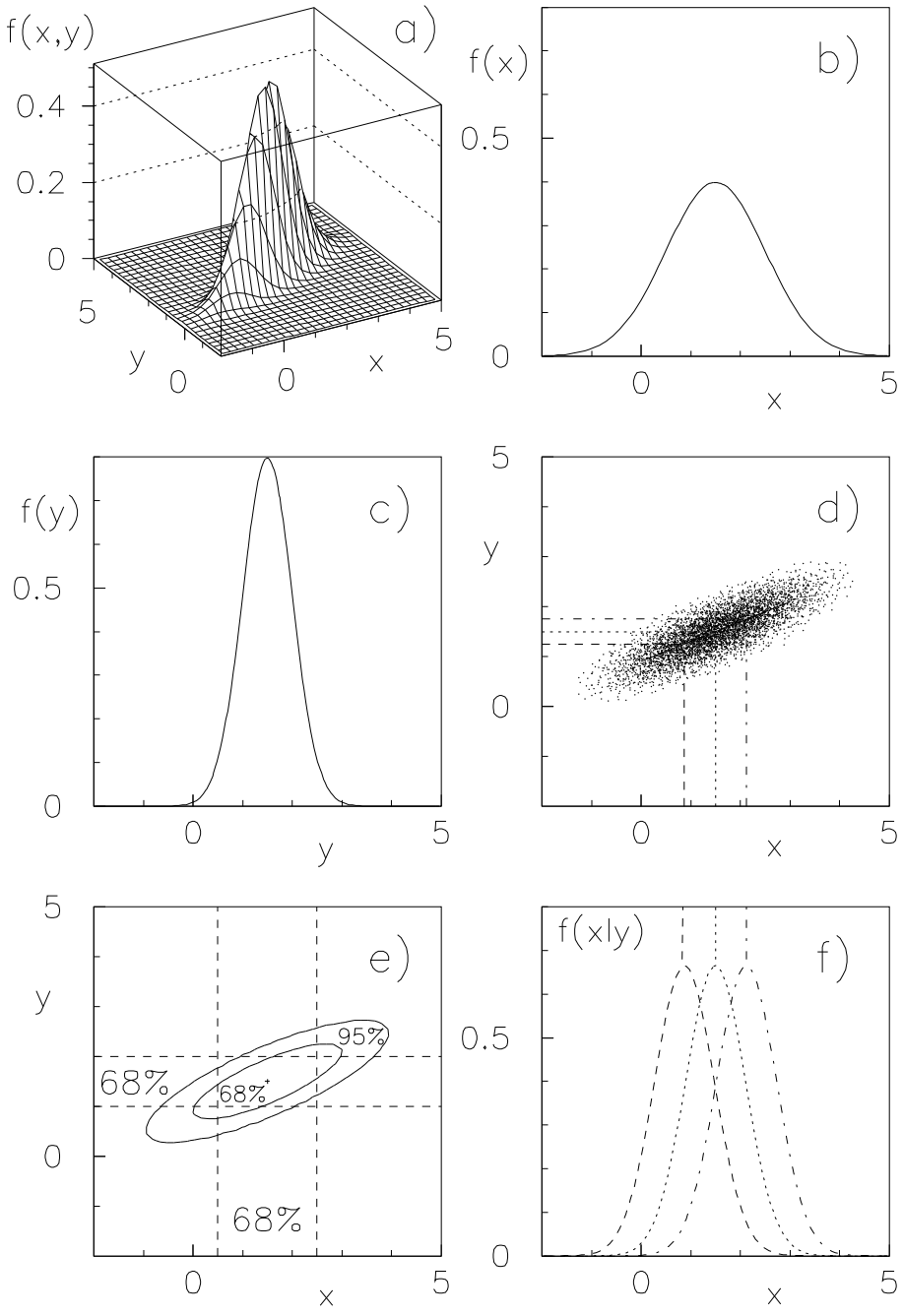

Figure 4.1: Example of bivariate normal distribution.

Bivariate normal distribution.

Joint probability density function of X and Y with correlation coefficient ρ (see Fig. 4.1):

$$f(x,y) = \frac{1}{2\,\pi\,\sigma_x\,\sigma_y\,\sqrt{1-\rho^2}}\cdot \exp\left\{-\frac{1}{2\,(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\,\rho\,\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\,\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}. \tag{4.67}$$

Marginal distributions:

$$X \sim \mathcal{N}(\mu_x,\sigma_x), \tag{4.68}$$

$$Y \sim \mathcal{N}(\mu_y,\sigma_y)\,. \tag{4.69}$$

Conditional distribution:

$$f(y\,|\,x_\circ) = \frac{1}{\sqrt{2\,\pi}\,\sigma_y\,\sqrt{1-\rho^2}}\exp\left[-\frac{\left(y-\left[\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}\,(x_\circ-\mu_x)\right]\right)^2}{2\,\sigma_y^2\,(1-\rho^2)}\right], \tag{4.70}$$

i.e.

$$Y_{|x_\circ} \sim \mathcal{N}\left(\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}\,(x_\circ-\mu_x)\,,\;\sigma_y\sqrt{1-\rho^2}\right). \tag{4.71}$$

The condition $X = x_\circ$ squeezes the standard deviation and shifts the mean of Y.

4.2 Central limit theorem

4.2.1 Terms and role

The well-known central limit theorem plays a crucial role in statistics and justifies the enormous importance that the normal distribution has in many practical applications (this is why it appears on 10 DM notes).

We have reminded ourselves in (4.61)–(4.62) of the expression of the mean and variance of a linear combination of random variables,

$$Y = \sum_{i=1}^{n} c_i X_i\,,$$

in the most general case, which includes correlated variables ($\rho_{ij} \neq 0$). In the case of independent variables the variance is given by the simpler, and better known, expression

$$\sigma_Y^2 = \sum_{i=1}^{n} c_i^2\,\sigma_i^2 \qquad (\rho_{ij}=0,\;i\neq j)\,. \tag{4.72}$$

This is a very general statement, valid for any number and kind of variables (with the obvious clause that all σ_i must be finite), but it does not give any information about the probability distribution of Y. Even if all X_i follow the same distributions $f(x)$, $f(y)$ is different from $f(x)$, with some exceptions, one of these being the normal.

The central limit theorem states that the distribution of a linear combination Y will be approximately normal if the variables X_i are independent and σ_Y^2 is much larger than any single component $c_i^2\sigma_i^2$ from a non-normally distributed X_i. The last condition is just to guarantee that there is no single random variable which dominates the fluctuations. The accuracy of the approximation improves as the number of variables n increases (the theorem says "when $n \to \infty$"):

$$n \to \infty \Longrightarrow Y \sim \mathcal{N}\left(\sum_{i=1}^{n} c_i \,\mathrm{E}(X_i), \left(\sum_{i=1}^{n} c_i^2\,\sigma_i^2\right)^{\frac{1}{2}}\right). \tag{4.73}$$

The proof of the theorem can be found in standard textbooks. For practical purposes, and if one is not very interested in the detailed behaviour of the tails, n equal to 2 or 3 may already give a satisfactory approximation, especially if the X_i exhibits a Gaussian-like shape. See, for example, Fig. 4.2, where samples of 10 000 events have been simulated, starting from a uniform distribution and from a crazy square-wave distribution. The latter, depicting a kind of worst practical case, shows that, already for $n = 20$ the distribution of the sum is practically normal. In the case of the uniform distribution $n = 3$ already gives an acceptable approximation as far as probability intervals of one or two standard deviations from the mean value are concerned. The figure also shows that, starting from a triangular distribution (obtained in the example from the sum of two uniform distributed variables), $n = 2$ is already sufficient (The sum of two triangular distributed variables is equivalent to the sum of four uniform distributed variables.)

4.2.2 Distribution of a sample average

As first application of the theorem, let us remind ourselves that a sample average $\overline{X}_n$ of n independent variables,

$$\overline{X}_n = \sum_{i=1}^{n} \frac{1}{n} X_i, \tag{4.74}$$

is normally distributed, since it is a linear combination of n variables X_i, with $c_i = 1/n$. Then,

$$\overline{X}_n \sim \mathcal{N}(\mu_{\overline{X}_n}, \sigma_{\overline{X}_n}), \tag{4.75}$$

$$\mu_{\overline{X}_n} = \sum_{i=1}^{n} \frac{1}{n}\,\mu = \mu, \tag{4.76}$$

$$\sigma^2_{\overline{X}_n} = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}, \tag{4.77}$$

$$\sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\,. \tag{4.78}$$

This result, we repeat, is independent of the distribution of X and is already approximately valid for small values of n.

4.2.3 Normal approximation of the binomial and of the Poisson distribution

Another important application of the theorem is that the binomial and the Poisson distribution can be approximated, for large numbers, by a normal distribution. This is a general result, valid for all distributions which have the reproductive property under the sum. Distributions of this kind are the binomial, the Poisson and the χ^2. Let us go into more detail:

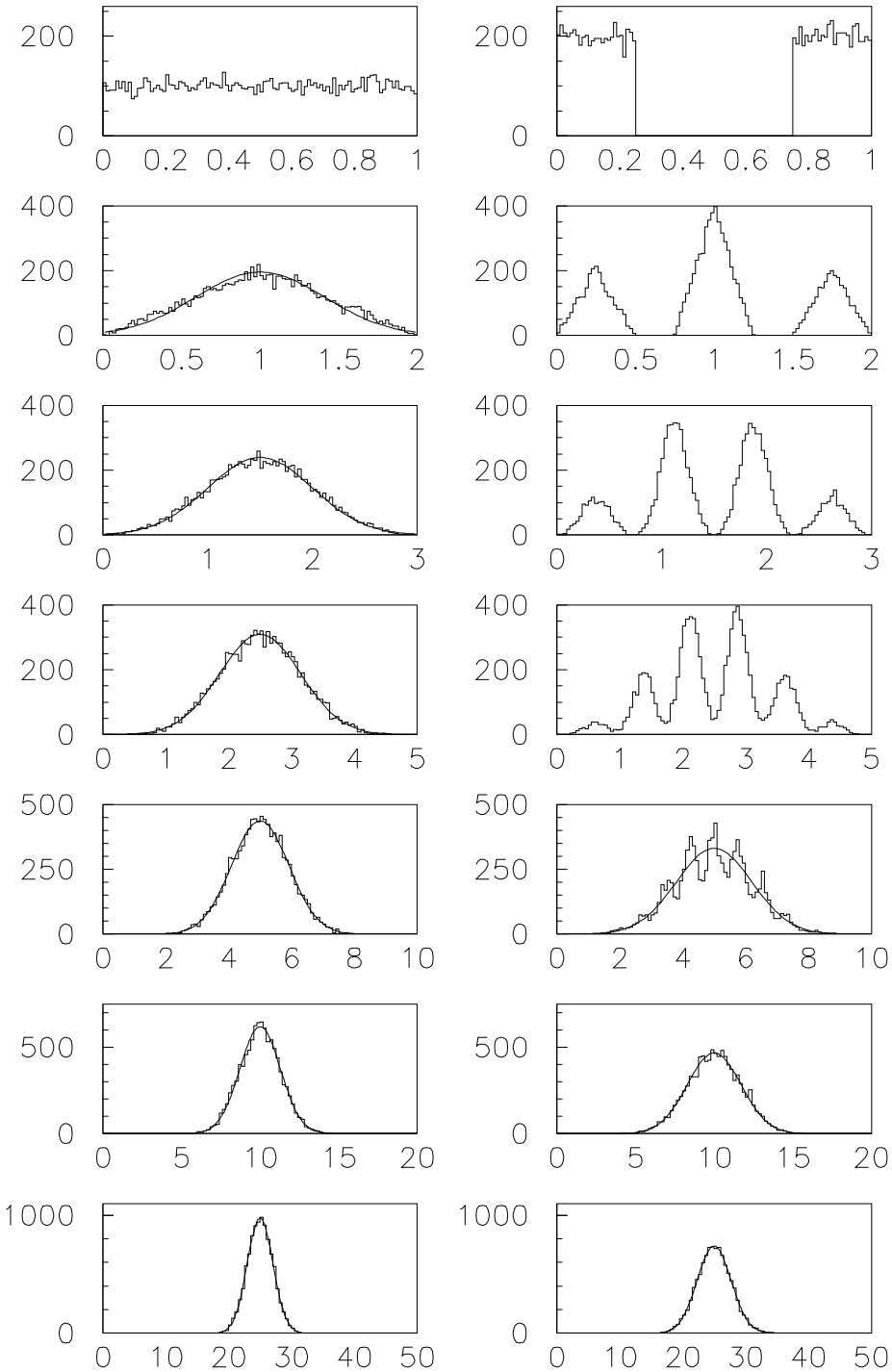

Figure 4.2: Central limit theorem at work: The sum of n variables, for two different distributions, is shown. The values of n (top to bottom) are 1, 2, 3, 5, 10, 20, 50.

$\boxed{\mathcal{B}_{n,p} \to \mathcal{N}\left(n\,p,\sqrt{n\,p\,(1-p)}\right)}$ The reproductive property of the binomial states that if X_1, X_2, $\ldots$, X_m are m independent variables, each following a binomial distribution of parameter n_i and p, then their sum $Y = \sum_i X_i$ also follows a binomial distribution with parameters $n = \sum_i n_i$ and p. It is easy to be convinced of this property without any mathematics. Just think of what happens if one tosses bunches of three, of five and of ten coins, and then one considers the global result: a binomial with a large n can then always be seen as a sum of many binomials with smaller n_i. The application of the central limit theorem is straightforward, apart from deciding when the convergence is acceptable. The parameters on which one has to base a judgment are in this case $\mu = n\,p$ and the complementary quantity $\mu^c = n\,(1-p) = n - \mu$. If they are both $\gtrsim 10$ then the approximation starts to be reasonable.

$\boxed{\mathcal{P}_\lambda \to \mathcal{N}\left(\lambda,\sqrt{\lambda}\right)}$ The same argument holds for the Poisson distribution. In this case the approximation starts to be reasonable when $\mu = \lambda \gtrsim 10$.

4.2.4 Normal distribution of measurement errors

The central limit theorem is also important to justify why in many cases the distribution followed by the measured values around their average is approximately normal. Often, in fact, the random experimental error e, which causes the fluctuations of the measured values around the unknown true value of the physical quantity, can be seen as an incoherent sum of smaller contributions e_i:

$$e = \sum_i e_i\,, \tag{4.79}$$

each contribution having a distribution which satisfies the conditions of the central limit theorem.

4.2.5 Caution

Following this commercial in favour of the miraculous properties of the central limit theorem, some words of caution are in order.

- Although I have tried to convince the reader that the convergence is rather fast in the cases of practical interest, the theorem only states that the asymptotic Gaussian distribution is reached for $n \to \infty$. As an example of very slow convergence, let us imagine 10^9 independent variables described by a Poisson distribution of $\lambda_i = 10^{-9}$: their sum is still far from a Gaussian.

- Sometimes the conditions of the theorem are not satisfied.

  - A single component dominates the fluctuation of the sum: a typical case is the well-known Landau distribution; systematic errors may also have the same effect on the global error.
  - The condition of independence is lost if systematic errors affect a set of measurements, or if there is coherent noise.

- The tails of the distributions do exist and they are not always Gaussian! Moreover, realizations of a random variable several standard deviations away from the mean are possible. And they show up without notice!

Chapter 5

Bayesian inference applied to measurements

" ... these problems are classified as probability of the causes, and are the most interesting of all from their scientific applications".

"An effect may be produced by the cause a or by the cause b. The effect has just been observed. We ask the probability that it is due to the cause a. This is an à posteriori probability of cause. But I could not calculate it, if a convention more or less justified did not tell me in advance what is the à priori probability for the cause a to come into play. I mean the probability of this event to some one who had not observed the effect."
(Henri Poincaré)

5.1 Measurement errors and measurement uncertainty

One might assume that the concepts of error and uncertainty are well enough known to be not worth discussing. Nevertheless a few comments are needed (although for more details the DIN [1] and ISO [3, 4] recommendations should be consulted).

- The first concerns the terminology. In fact the words error and uncertainty are currently used almost as synonyms:
  - 'error' to mean both error and uncertainty (but nobody says 'Heisenberg Error Principle');
  - 'uncertainty' only for the uncertainty.

  'Usually' we understand what each is talking about, but a more precise use of these nouns would really help. This is strongly called for by the DIN [1] and ISO [3, 4] recommendations. They state in fact that
  - error is *"the result of a measurement minus a true value of the measurand"* – it follows that the error is usually unkown;
  - uncertainty is a *"parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand"*;

- Within the HEP community there is an established practice for reporting the final uncertainty of a measurement in the form of standard deviation. This is also recommended by the mentioned standards. However, this should be done at each step of the analysis, instead of estimating maximum error bounds and using them as standard deviation in the error propagation.

- The process of measurement is a complex one and it is difficult to disentangle the different contributions which cause the total error. In particular, the active role of the experimentalist is sometimes overlooked. For this reason it is often incorrect to quote the (nominal) uncertainty due to the instrument as if it were the uncertainty of the measurement.

5.2 Statistical inference

5.2.1 Bayesian inference

In the Bayesian framework the inference is performed by calculating the final distribution of the random variable associated with the true values of the physical quantities from all available information. Let us call $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ the n-tuple (vector) of observables, $\underline{\mu} = \{\mu_1, \mu_2, \ldots, \mu_n\}$ the n-tuple of the true values of the physical quantities of interest, and $\underline{h} = \{h_1, h_2, \ldots, h_n\}$ the n-tuple of all the possible realizations of the influence variables H_i. The term "influence variable" is used here with an extended meaning, to indicate not only external factors which could influence the result (temperature, atmospheric pressure, and so on) but also any possible calibration constant and any source of systematic errors. In fact the distinction between $\underline{\mu}$ and $\underline{h}$ is artificial, since they are all conditional hypotheses. We separate them simply because at the end we will marginalize the final joint distribution functions with respect to $\underline{\mu}$, integrating the joint distribution with respect to the other hypotheses considered as influence variables.

The likelihood of the sample $\underline{x}$ being produced from $\underline{h}$ and $\underline{\mu}$ and the initial probability are

$$f(\underline{x} \,|\, \underline{\mu}, \underline{h}, H_\circ)$$

and

$$f_\circ(\underline{\mu}, \underline{h}) = f(\underline{\mu}, \underline{h} \,|\, H_\circ)\,, \tag{5.1}$$

respectively. $H_\circ$ is intended to remind us, yet again, that likelihoods and priors — and hence conclusions — depend on all explicit and implicit assumptions within the problem, and in particular on the parametric functions used to model priors and likelihoods. To simplify the formulae, $H_\circ$ will no longer be written explicitly.

Using the Bayes formula for multidimensional continuous distributions [an extension of (4.52)] we obtain the most general formula of inference,

$$f(\underline{\mu}, \underline{h} \,|\, \underline{x}) = \frac{f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})}{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{\mu}\, \mathrm{d}\underline{h}}\,, \tag{5.2}$$

yielding the joint distribution of all conditional variables $\underline{\mu}$ and $\underline{h}$ which are responsible for the observed sample $\underline{x}$. To obtain the final distribution of $\underline{\mu}$ one has to integrate (5.2) over all possible values of $\underline{h}$, obtaining

$$\boxed{f(\underline{\mu} \,|\, \underline{x}) = \frac{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{h}}{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{\mu}\, \mathrm{d}\underline{h}}\,.} \tag{5.3}$$

Apart from the technical problem of evaluating the integrals, if need be numerically or using Monte Carlo methods,[1] (5.3) represents the most general form of hypothetical inductive inference. The word 'hypothetical' reminds us of $H_\circ$.

When all the sources of influence are under control, i.e. they can be assumed to take a precise value, the initial distribution can be factorized by a $f_\circ(\underline{\mu})$ and a Dirac $\delta(\underline{h} - \underline{h}_\circ)$, obtaining the much simpler formula

$$\begin{aligned} f(\underline{\mu} \,|\, \underline{x}) &= \frac{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu})\, \delta(\underline{h} - \underline{h}_\circ)\, \mathrm{d}\underline{h}}{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h})\, f_\circ(\underline{\mu})\, \delta(\underline{h} - \underline{h}_\circ)\, \mathrm{d}\underline{\mu}\, \mathrm{d}\underline{h}} \\ &= \frac{f(\underline{x} \,|\, \underline{\mu}, \underline{h}_\circ)\, f_\circ(\underline{\mu})}{\int f(\underline{x} \,|\, \underline{\mu}, \underline{h}_\circ)\, f_\circ(\underline{\mu})\, \mathrm{d}\underline{\mu}}\,. \end{aligned} \tag{5.4}$$

Even if formulae (5.3)–(5.4) look complicated because of the multidimensional integration and of the continuous nature of $\underline{\mu}$, conceptually they are identical to the example of the $\mathrm{d}E/\mathrm{d}x$ measurement discussed in Section 3.4.3.

The final probability density function provides the most complete and detailed information about the unknown quantities, but sometimes (almost always ...) one is not interested in full knowledge of $f(\underline{\mu})$, but just in a few numbers which summarize at best the position and the width of the distribution (for example when publishing the result in a journal in the most compact way). The most natural quantities for this purpose are the expected value and the variance, or the standard deviation. Then the Bayesian best estimate of a physical quantity is:

$$\widehat{\mu}_i = \mathrm{E}[\mu_i] = \int \mu_i\, f(\underline{\mu} \,|\, \underline{x})\, \mathrm{d}\underline{\mu}, \tag{5.5}$$

$$\sigma^2_{\mu_i} \equiv \mathrm{Var}(\mu_i) = \mathrm{E}[\mu_i^2] - \mathrm{E}^2[\mu_i]. \tag{5.6}$$

When many true values are inferred from the same data the numbers which synthesize the result are not only the expected values and variances, but also the covariances, which give at least the correlation coefficients between the variables:

$$\rho_{ij} \equiv \rho(\mu_i, \mu_j) = \frac{\mathrm{Cov}(\mu_i, \mu_j)}{\sigma_{\mu_i}\, \sigma_{\mu_j}}\,. \tag{5.7}$$

In the following sections we will deal in most cases with only one value to infer:

$$f(\mu \,|\, \underline{x}) = \ldots\,. \tag{5.8}$$

5.2.2 Bayesian inference and maximum likelihood

We have already said that the dependence of the final probabilities on the initial ones gets weaker as the amount of experimental information increases. Without going into mathematical complications (the proof of this statement can be found for example in Ref. [29]) this simply means that, asymptotically, whatever $f_\circ(\mu)$ one puts in (5.4), $f(\mu \,|\, \underline{x})$ is unaffected. This happens when the width of $f_\circ(\mu)$ is much larger than that of the likelihood, when the latter is considered as a mathematical function of μ. Therefore $f_\circ(\mu)$ acts as a constant in the region of μ where the likelihood is significantly different from 0. This is equivalent to dropping $f_\circ(\mu)$ from (5.4). This results in

$$f(\mu \,|\, \underline{x}) \approx \frac{f(\underline{x} \,|\, \mu, \underline{h}_\circ)}{\int f(\underline{x} \,|\, \mu, \underline{h}_\circ)\, \mathrm{d}\mu}\,. \tag{5.9}$$

[1] This is conceptually what experimentalists do when they change all the parameters of the Monte Carlo simulation in order to estimate the systematic error.

Since the denominator of the Bayes formula has the technical role of properly normalizing the probability density function, the result can be written in the simple form

$$f(\mu \,|\, \underline{x}) \propto f(\underline{x} \,|\, \mu, \underline{h}_\circ) \text{“} \equiv \mathcal{L}(\mu; \underline{x}, \underline{h}_\circ) \text{”} \,. \tag{5.10}$$

Asymptotically the final probability is just the (normalized) likelihood! The notation $\mathcal{L}$ is that used in the maximum likelihood literature (note that, not only does f become $\mathcal{L}$, but also “|” has been replaced by “;”: $\mathcal{L}$ has no probabilistic interpretation, when referring to μ, in conventional statistics.)

If the mean value of $f(\mu \,|\, \underline{x})$ coincides with the value for which $f(\mu \,|\, \underline{x})$ has a maximum, we obtain the maximum likelihood method. This does not mean that the Bayesian methods are ‘blessed’ because of this achievement, and hence they can be used only in those cases where they provide the same results. It is the other way round: The maximum likelihood method gets justified when all the limiting conditions of the approach ($\rightarrow$ insensitivity of the result from the initial probability $\rightarrow$ large number of events) are satisfied.

Even if in this asymptotic limit the two approaches yield the same numerical results, there are differences in their interpretation:

- The likelihood, after proper normalization, has a probabilistic meaning for Bayesians but not for frequentists; so Bayesians can say that the probability that μ is in a certain interval is, for example, 68%, while this statement is blasphemous for a frequentist (the true value is a constant from his point of view).

- Frequentists prefer to choose $\widehat{\mu}_L$, the value which maximizes the likelihood, as estimator. For Bayesians, on the other hand, the expected value $\widehat{\mu}_B = \mathrm{E}[\mu]$ (also called the prevision) is more appropriate. This is justified by the fact that the assumption of the $\mathrm{E}[\mu]$ as best estimate of μ minimizes the risk of a bet (always keep the bet in mind!). For example, if the final distribution is exponential with parameter τ (let us think for a moment of particle decays) the maximum likelihood method would recommend betting on the value $t = 0$, whereas the Bayesian approach suggests the value $t = \tau$. If the terms of the bet are ‘whoever gets closest wins’, what is the best strategy? And then, what is the best strategy if the terms are ‘whoever gets the exact value wins’? But now think of the probability of getting the exact value and of the probability of getting closest.

5.2.3 The dog, the hunter and the biased Bayesian estimators

One of the most important tests to judge the quality of an estimator is whether or not it is correct (not biased). Maximum likelihood estimators are usually correct, while Bayesian estimators — analysed within the maximum likelihood framework — often are not. This could be considered a weak point; however the Bayes estimators are simply naturally consistent with the state of information before new data become available. In the maximum likelihood method, on the other hand, it is not clear what the assumptions are.

Let us take an example which shows the logic of frequentistic inference and why the use of reasonable prior distributions yields results which that frame classifies as distorted. Imagine meeting a hunting dog in the country. Let us assume we know that there is a 50% probability of finding the dog within a radius of 100 m centred on the position of the hunter (this is our likelihood). Where is the hunter? He is with 50% probability within a radius of 100 m around the position of the dog, with equal probability in all directions: Obvious! This is exactly the logic scheme used in the frequentistic approach to build confidence regions from the estimator (the dog in this example). This however assumes that the hunter can be anywhere in the country.

But now let us change the state of information: the dog is by a river; the dog has collected a duck and runs in a certain direction; the dog is sleeping; the dog is in a field surrounded by a fence through which he can pass without problems, but the hunter cannot. Given any new condition the conclusion changes. Some of the new conditions change our likelihood, but some others only influence the initial distribution. For example, the case of the dog in an enclosure inaccessible to the hunter is exactly the problem encountered when measuring a quantity close to the edge of its physical region, which is quite common in frontier research.

5.3 Choice of the initial probability density function

5.3.1 Difference with respect to the discrete case

The title of this section is similar to that of Section 3.6, but the problem and the conclusions will be different. There we said that the Indifference Principle (or, in its refined modern version, the Maximum Entropy Principle) was a good choice. Here there are problems with infinities and with the fact that it is possible to map an infinite number of points contained in a finite region onto an infinite number of points contained in a larger or smaller finite region. This changes the probability density function. If, moreover, the transformation from one set of variables to the other is not linear (see, e.g., Fig. 5.1), what is uniform in one variable (X) is not uniform in another variable (e.g. $Y = X^2$). This problem does not exist in the case of discrete variables, since if $X = x_i$ has a probability $f(x_i)$ then $Y = x_i^2$ has the same probability. A different way of stating the problem is that the Jacobian of the transformation squeezes or stretches the metrics, changing the probability density function.

We will not enter into the open discussion about the optimal choice of the distribution. Essentially we shall use the uniform distribution, being careful to employ the variable which seems most appropriate for the problem, but you may disagree — surely with good reason — if you have a different kind of experiment in mind.

The same problem is also present, but well hidden, in the maximum likelihood method. For example, it is possible to demonstrate that, in the case of normally distributed likelihoods, a uniform distribution of the mean μ is implicitly assumed (see Section 5.4). There is nothing wrong with this, but one should be aware of it.

5.3.2 Bertrand paradox and angels' sex

A good example to help understand the problems outlined in the previous section is the so-called Bertrand paradox.

Problem: Given a circle of radius R and a chord drawn randomly on it, what is the probability that the length L of the chord is smaller than R?

Solution 1: Choose randomly two points on the circumference and draw a chord between them: $\Rightarrow P(L < R) = 1/3 = 0.33$.

Solution 2: Choose a straight line passing through the centre of the circle; then draw a second line, orthogonal to the first, and which intersects it inside the circle at a random distance from the centre: $\Rightarrow P(L < R) = 1 - \sqrt{3}/2 = 0.13$.

Solution 3: Choose randomly a point inside the circle and draw a straight line orthogonal to the radius that passes through the chosen point $\Rightarrow P(L < R) = 1/4 = 0.25$.

Your solution: ... ...?

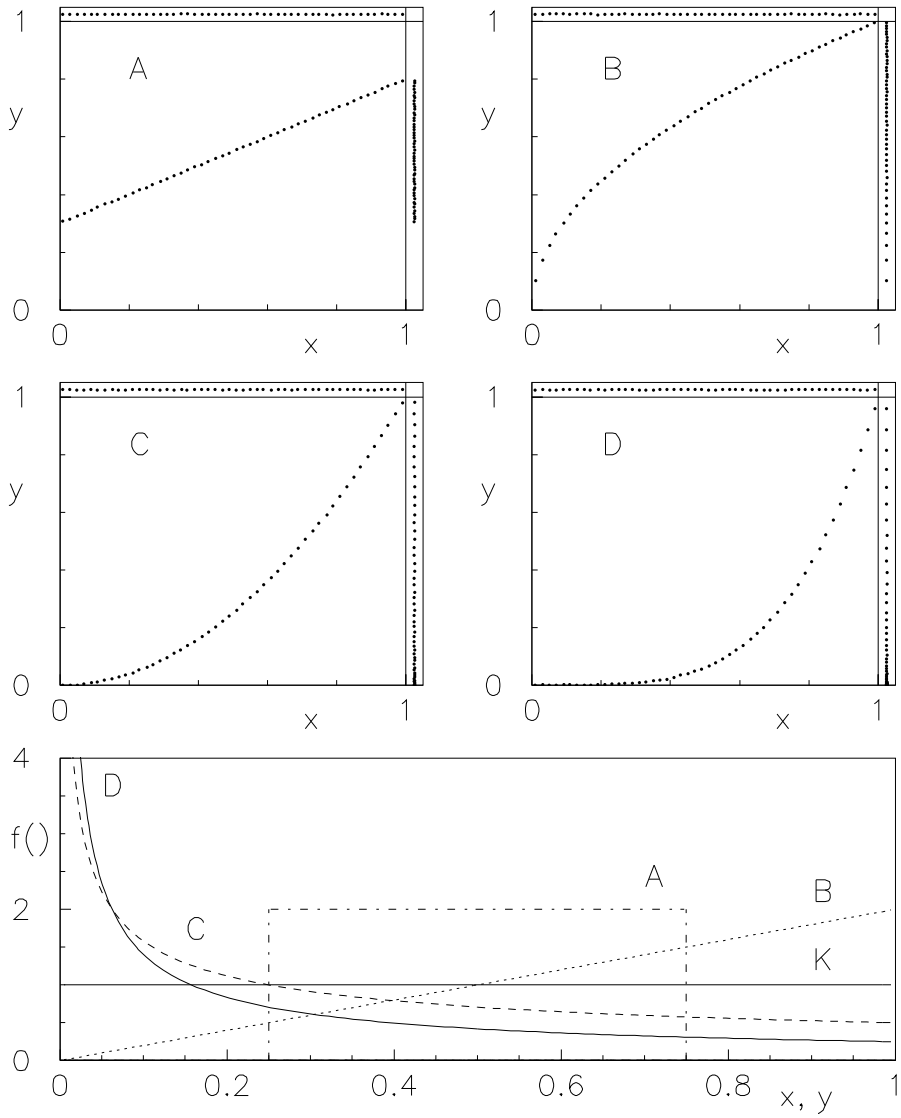

Figure 5.1: Examples of variable changes starting from a uniform distribution (K): A) $Y = 0.5\,X + 0.25$; B) $Y = \sqrt{X}$; C) $Y = X^2$; D) $Y = X^4$.

Question: What is the origin of the paradox?

Answer: The problem does not specify how to randomly choose the chord. The three solutions take a uniform distribution: along the circumference; along the radius; and inside the circle. What is uniform in one variable is not uniform in the others!

Question: Which is the right solution?

In principle you may imagine an infinite number of different solutions. From a physicist's viewpoint any attempt to answer this question is a waste of time. The reason why the paradox has been compared to the Byzantine discussions about the sex of angels is that there are indeed people arguing about it. For example, there is a school of thought which insists that Solution 2 is the right one.

In fact this kind of paradox, together with abuse of the Indifference Principle to assess, for example, the probability that the sun will rise tomorrow morning, threw a shadow over Bayesian methods at the end of the last century. The maximum likelihood method, which does not make explicit use of prior distributions, was then seen as a valid solution to the problem. But in reality the ambiguity of the proper metrics on which the initial distribution is uniform has an equivalent in the arbitrariness of the variable used in the likelihood function. In the end, what was criticized when it was stated explicitly in the Bayes formula is accepted passively when it is hidden in the maximum likelihood method.

5.4 Normally distributed observables

5.4.1 Final distribution, prevision and credibility intervals of the true value

The first application of the Bayesian inference will be that of a normally distributed quantity. Let us take a data sample $\underline{q}$ of n_1 measurements, of which we calculate the average $\overline{q}_{n_1}$. In our formalism $\overline{q}_{n_1}$ is a realization of the random variable $\overline{Q}_{n_1}$. Let us assume we know the standard deviation σ of the variable Q, either because n_1 is very large and can be estimated accurately from the sample or because it was known *a priori* (We are not going to discuss in these notes the case of small samples and unknown variance.)[2] The property of the average (see Section 4.2.2) tells us that the likelihood $f(\overline{Q}_{n_1} \,|\, \mu, \sigma)$ is Gaussian:

$$\overline{Q}_{n_1} \sim \mathcal{N}(\mu, \sigma/\sqrt{n_1}). \tag{5.11}$$

To simplify the following notation, let us call x_1 this average and σ_1 the standard deviation of the average:

$$x_1 = \overline{q}_{n_1}, \tag{5.12}$$

$$\sigma_1 = \sigma/\sqrt{n_1}\,. \tag{5.13}$$

We then apply (5.4) and get

$$f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1)) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)}{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)\,\mathrm{d}\mu}\,. \tag{5.14}$$

At this point we have to make a choice for $f_\circ(\mu)$. A reasonable choice is to take, as a first guess, a uniform distribution defined over a large interval which includes x_1. It is not really important how large the interval is, for a few σ_1 away from x_1 the integrand at the denominator tends to zero because of the Gaussian function. What is important is that a constant $f_\circ(\mu)$ can be simplified in (5.14), obtaining

$$f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1)) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\,\mathrm{d}\mu}\,. \tag{5.15}$$

The integral in the denominator is equal to unity, since integrating with respect to μ is equivalent to integrating with respect to x_1. The final result is then

$$f(\mu) = f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1)) = \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(\mu-x_1)^2}{2\,\sigma_1^2}}\quad : \tag{5.16}$$

[2] **Note added:** for criticisms about the standard treatment of the small-sample problem, see Ref. [22].

- the true value is normally distributed around x_1;
- its best estimate (prevision) is $\mathrm{E}[\mu] = x_1$;
- its variance is $\sigma_\mu = \sigma_1$;
- the confidence intervals, or 'credibility intervals', in which there is a certain probability of finding the true value are easily calculable:

Probability level (confidence level) (%)	Credibility interval (confidence interval)
68.3	$x_1 \pm \sigma_1$
90.0	$x_1 \pm 1.65\sigma_1$
95.0	$x_1 \pm 1.96\sigma_1$
99.0	$x_1 \pm 2.58\sigma_1$
99.73	$x_1 \pm 3\sigma_1$

5.4.2 Combination of several measurements

Let us imagine making a second set of measurements of the physical quantity, which we assume unchanged from the previous set of measurements. How will our knowledge of μ change after this new information? Let us call $x_2 = \overline{q}_{n_2}$ and $\sigma_2 = \sigma'/\sqrt{n_2}$ the new average and standard deviation of the average (σ' may be different from σ of the sample of n_1 measurements), respectively. Applying Bayes' theorem a second time we now have to use as initial distribution the final probability of the previous inference:

$$f(\mu \,|\, x_1, \sigma_1, x_2, \sigma_2, \mathcal{N}) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_2}\, e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\, f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1))}{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_2}\, e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\, f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1))\, \mathrm{d}\mu}\,. \tag{5.17}$$

The integral is not as simple as the previous one, but still feasible analytically. The final result is

$$f(\mu \,|\, x_1, \sigma_1, x_2, \sigma_2, \mathcal{N}) = \frac{1}{\sqrt{2\,\pi}\,\sigma_A}\, e^{-\frac{(\mu-x_A)^2}{2\,\sigma_A^2}}\,, \tag{5.18}$$

where

$$x_A = \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\,, \tag{5.19}$$

$$\frac{1}{\sigma_A^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,. \tag{5.20}$$

One recognizes the famous formula of the weighted average with the inverse of the variances, usually obtained from maximum likelihood. There are some comments to be made.

- Bayes' theorem updates the knowledge about μ in an automatic and natural way.

- If $\sigma_1 \gg \sigma_2$ (and x_1 is not too far from x_2) the final result is only determined by the second sample of measurements. This suggests that an alternative vague *a priori* distribution can be, instead of uniform, a Gaussian with a large enough variance and a reasonable mean.

- The combination of the samples requires a subjective judgement that the two samples are really coming from the same true value μ. We will not discuss this point in these notes,[3] but a hint on how to proceed is to take the inference on the difference of two measurements, D, as explained at the end of Section 5.6.1 and judge yourself whether $D = 0$ is consistent with the probability density function of D.

5.4.3 Measurements close to the edge of the physical region

A case which has essentially no solution in the maximum likelihood approach is when a measurement is performed at the edge of the physical region and the measured value comes out very close to it, or even on the unphysical region. Let us take a numerical example.

Problem: An experiment is planned to measure the (electron) neutrino mass. The simulations show that the mass resolution is $3.3\,\mathrm{eV}/c^2$, largely independent of the mass value, and that the measured mass is normally distributed around the true mass.[4] The mass value which results from the analysis procedure,[5] and corrected for all known systematic effects, is $x = -5.41\,\mathrm{eV}/c^2$. What have we learned about the neutrino mass?

Solution: Our *a priori* value of the mass is that it is positive and not too large (otherwise it would already have been measured in other experiments). One can take any vague distribution which assigns a probability density function between 0 and 20 or 30 eV/c^2. In fact, if an experiment having a resolution of $\sigma = 3.3\,\mathrm{eV}/c^2$ has been planned and financed by rational people, with the hope of finding evidence of non-negligible mass, it means that the mass was thought to be in that range. If there is no reason to prefer one of the values in that interval a uniform distribution can be used, for example

$$f_{\circ K}(m) = k = 1/30 \qquad (0 \le m \le 30)\,. \tag{5.21}$$

Otherwise, if one thinks there is a greater chance of the mass having small rather than high values, a prior which reflects such an assumption could be chosen, for example a half normal with $\sigma_\circ = 10\,\mathrm{eV}$

$$f_{\circ N}(m) = \frac{2}{\sqrt{2\,\pi}\,\sigma_\circ}\,\exp\left[-\frac{m^2}{2\,\sigma_\circ^2}\right] \qquad (m \ge 0)\,, \tag{5.22}$$

or a triangular distribution

$$f_{\circ T}(m) = \frac{1}{450}\,(30 - x) \qquad (0 \le m \le 30)\,. \tag{5.23}$$

[3] **Note added**: as is easy to imagine, the problem of the 'outliers' should be treated with care, and surely avoiding automatic prescriptions. Some hints can be found in Refs. [43] and [44], and references therein.

[4] In reality, often m^2 rather than m is normally distributed. In this case the terms of the problem change and a new solution should be worked out, following the trace indicated in this example.

[5] We consider detector and analysis machinery as a black box, no matter how complicated it is, and treat the numerical outcome as a result of a direct measurement [1].

Let us consider for simplicity the uniform distribution

$$f(m\,|\,x, f_{\circ K}) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right] k}{\int_0^{30}\frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right] k\,\mathrm{d}\mu} \tag{5.24}$$

$$= \frac{19.8}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(m-x)^2}{2\sigma^2}\right] \qquad (0 \le m \le 30)\,. \tag{5.25}$$

The value which has the highest degree of belief is $m = 0$, but $f(m)$ is non-vanishing up to $30\,\mathrm{eV}/c^2$ (even if very small). We can define an interval, starting from $m = 0$, in which we believe that m should have a certain probability. For example this level of probability can be 95%. One has to find the value $m_\circ$ for which the cumulative function $F(m_\circ)$ equals 0.95. This value of m is called the upper limit (or upper bound). The result is

$$m < 3.9\,\mathrm{eV}/c^2 \quad \text{at } 0.95\% \text{ probability}\,. \tag{5.26}$$

If we had assumed the other initial distributions the limit would have been in both cases

$$m < 3.7\,\mathrm{eV}/c^2 \quad \text{at } 0.95\% \text{ probability}\,, \tag{5.27}$$

practically the same (especially if compared with the experimental resolution of $3.3\,\mathrm{eV}/c^2$).

Comment: Let us assume an *a priori* function sharply peaked at zero and see what happens. For example it could be of the kind

$$f_{\circ S}(m) \propto \frac{1}{m}\,. \tag{5.28}$$

To avoid singularities in the integral, let us take a power of m slightly greater than -1, for example -0.99, and let us limit its domain to 30, getting

$$f_{\circ S}(m) = \frac{0.01 \cdot 30^{0.01}}{m^{0.99}}\,. \tag{5.29}$$

The upper limit becomes

$$m < 0.006\,\mathrm{eV}/c^2 \quad \text{at } 0.95\% \text{ probability}\,. \tag{5.30}$$

Any experienced physicist would find this result ridiculous. The upper limit is less than 0.2% of the experimental resolution; rather like expecting to resolve objects having dimensions smaller than a micron with a design ruler! Note instead that in the previous examples the limit was always of the order of magnitude of the experimental resolution σ. As $f_{\circ S}(m)$ becomes more and more peaked at zero (power of $x \to 1$) the limit gets smaller and smaller. This means that, asymptotically, the degree of belief that $m = 0$ is so high that whatever you measure you will conclude that $m = 0$: you could use the measurement to calibrate the apparatus! This means that this choice of initial distribution was unreasonable.

Instead, priors motivated by the positive attitude of the researchers are much more stable, and even when the observation is very negative the result is stable, and one always gets a limit of the order of the experimental resolution. Anyhow, it is also clear that when x is several σ below zero one starts to suspect that something is wrong with the experiment, which formally corresponds to doubts about the likelihood itself.

5.5 Counting experiments

5.5.1 Binomially distributed observables

Let us assume we have performed n trials and obtained x favourable events. What is the probability of the next event? This situation happens frequently when measuring efficiencies, branching ratios, etc. Stated more generally, one tries to infer the constant and unknown probability[6] of an event occurring.

Where we can assume that the probability is constant and the observed number of favourable events are binomially distributed, the unknown quantity to be measured is the parameter p of the binomial. Using Bayes' theorem we get

$$\begin{aligned} f(p \,|\, x, n, \mathcal{B}) &= \frac{f(x \,|\, \mathcal{B}_{n,p})\, f_\circ(p)}{\int_0^1 f(x \,|\, \mathcal{B}_{n,p})\, f_\circ(p)\, \mathrm{d}p} \\ &= \frac{\frac{n!}{(n-x)!\,x!}\, p^x\, (1-p)^{n-x}\, f_\circ(p)}{\int_0^1 \frac{n!}{(n-x)!\,x!}\, p^x\, (1-p)^{n-x}\, f_\circ(p)\, \mathrm{d}p} \\ &= \frac{p^x\, (1-p)^{n-x}}{\int_0^1 p^x\, (1-p)^{n-x}\, \mathrm{d}p}\,, \end{aligned} \tag{5.31}$$

where an initial uniform distribution has been assumed. The final distribution is known to statisticians as Beta distribution since the integral at the denominator is the special function called β, defined also for real values of x and n (technically this is a Beta with parameters $a = x + 1$ and $b = n - x + 1$). In our case these two numbers are integer and the integral becomes equal to $x!\,(n-x)!/(n+1)!$. We then get

$$f(p \,|\, x, n, \mathcal{B}) = \frac{(n+1)!}{x!\,(n-x)!}\, p^x\, (1-p)^{n-x}\,, \tag{5.32}$$

some examples of which are shown in Fig. 5.2.

Expected value and the variance of this distribution are:

[6] This concept, which is very close to the physicist's mentality, is not correct from the probabilistic — cognitive — point of view. According to the Bayesian scheme, in fact, the probability changes with the new observations. The final inference of p, however, does not depend on the particular sequence yielding x successes over n trials. This can be seen in the next table where $f_n(p)$ is given as a function of the number of trials n, for the three sequences which give two successes (indicated by 1) in three trials [the use of (5.32) is anticipated]:

	Sequence		
n	011	101	110
0	1	1	1
1	$2\,(1-p)$	$2\,p$	$2\,p$
2	$6\,p\,(1-p)$	$6\,p\,(1-p)$	$3\,p^2$
3	$12\,p^2\,(1-p)$	$12\,p^2\,(1-p)$	$12\,p^2\,(1-p)$

This important result, related to the concept of exchangeability, allows a physicist who is reluctant to give up the concept of unknown constant probability to see the problem from his point of view, ensuring that the same numerical result is obtained.

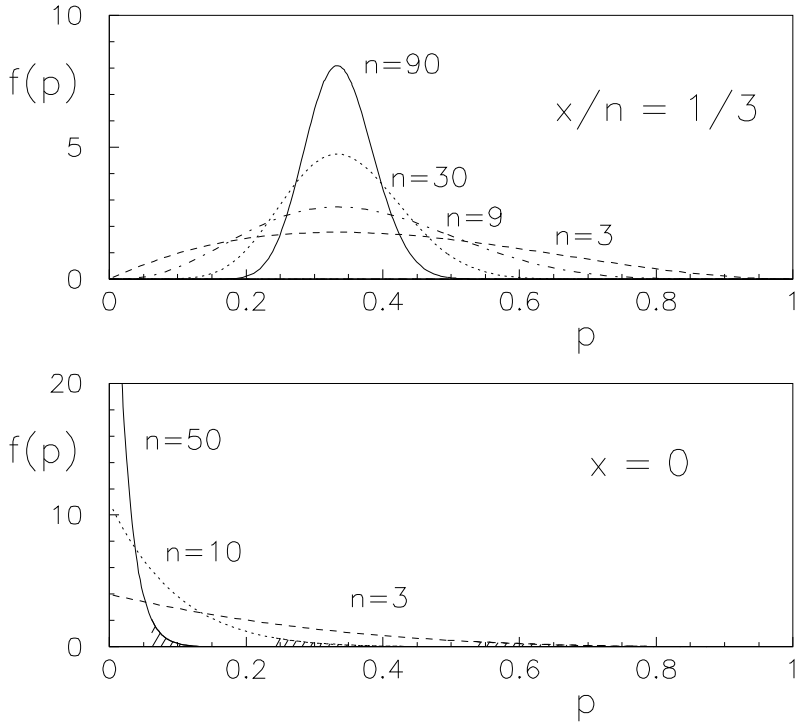

Figure 5.2: Probability density function of the binomial parameter p, having observed x successes in n trials.

$$\mathrm{E}[p] = \frac{x+1}{n+2}, \tag{5.33}$$

$$\begin{aligned} \mathrm{Var}(p) &= \frac{(x+1)(n-x+1)}{(n+3)(n+2)^2} \qquad (5.34)\\ &= \frac{x+1}{n+2}\left(\frac{n+2}{n+2} - \frac{x+1}{n+2}\right)\frac{1}{n+3} \\ &= \mathrm{E}[p]\ (1-\mathrm{E}[p])\ \frac{1}{n+3}\,. \qquad (5.35) \end{aligned}$$

The value of p for which $f(p)$ has the maximum is instead $p_m = x/n$. The expression $\mathrm{E}[p]$ gives the prevision of the probability for the $(n+1)$-th event occurring and is called the recursive Laplace formula, or Laplace's rule of succession.

When x and n become large, and $0 \ll x \ll n$, $f(p)$ has the following asymptotic properties:

$$\mathrm{E}[p] \approx p_m = \frac{x}{n}, \tag{5.36}$$

$$\mathrm{Var}(p) \approx \frac{x}{n}\left(1-\frac{x}{n}\right)\frac{1}{n} = \frac{p_m\,(1-p_m)}{n}, \tag{5.37}$$

$$\sigma_p \approx \sqrt{\frac{p_m\,(1-p_m)}{n}}, \tag{5.38}$$

$$p \sim \mathcal{N}(p_m, \sigma_p)\,. \tag{5.39}$$

Under these conditions the frequentistic definition (evaluation rule!) of probability (x/n) is recovered.

Let us see two particular situations: when $x = 0$ and $x = n$. In these cases one gives the result as upper or lower limits, respectively. Let us sketch the solutions:

- $\underline{x = n}$:

$$f(n \,|\, \mathcal{B}_{n,p}) = p^n, \tag{5.40}$$

$$f(p \,|\, x = n, \mathcal{B}) = \frac{p^n}{\int_0^1 p^n \,\mathrm{d}p} = (n+1)\, p^n, \tag{5.41}$$

$$F(p \,|\, x = n, \mathcal{B}) = p^{n+1}\,. \tag{5.42}$$

To get the 95% lower bound (limit):

$$F(p_\circ \,|\, x = n, \mathcal{B}) = 0.05\,,$$

$$p_\circ = \sqrt[n+1]{0.05}\,. \tag{5.43}$$

An increasing number of trials n constrain more and more p around 1.

- $\underline{x = 0}$:

$$f(0 \,|\, \mathcal{B}_{n,p}) = (1-p)^n, \tag{5.44}$$

$$f(p \,|\, x = 0, n, \mathcal{B}) = \frac{(1-p)^n}{\int_0^1 (1-p)^n \,\mathrm{d}p} = (n+1)\,(1-p)^n, \tag{5.45}$$

$$F(p \,|\, x = 0, n, \mathcal{B}) = 1 - (1-p)^{n+1}\,. \tag{5.46}$$

To get the 95% upper bound (limit):

$$F(p_\circ \,|\, x = 0, n, \mathcal{B}) = 0.95,$$

$$p_\circ = 1 - \sqrt[n+1]{0.05}\,. \tag{5.47}$$

The following table shows the 95% probability limits as a function of n. The Poisson approximation, to be discussed in the next section, is also shown.

	Probability level = 95%		
n	$x = n$	$x = 0$	
	binomial	binomial	Poisson approx. ($p_\circ = 3/n$)
3	$p \geq 0.47$	$p \leq 0.53$	$p \leq 1$
5	$p \geq 0.61$	$p \leq 0.39$	$p \leq 0.6$
10	$p \geq 0.76$	$p \leq 0.24$	$p \leq 0.3$
50	$p \geq 0.94$	$p \leq 0.057$	$p \leq 0.06$
100	$p \geq 0.97$	$p \leq 0.029$	$p \leq 0.03$
1000	$p \geq 0.997$	$p \leq 0.003$	$p \leq 0.003$

To show in this simple case how $f(p)$ is updated by the new information, let us imagine we have performed two experiments. The results are $x_1 = n_1$ and $x_2 = n_2$, respectively. Obviously the global information is equivalent to $x = x_1 + x_2$ and $n = n_1 + n_2$, with $x = n$. We then get

$$f(p \,|\, x = n, \mathcal{B}) = (n+1)\, p^n = (n_1 + n_2 + 1)\, p^{n_1+n_2}\,. \tag{5.48}$$

A different way of proceeding would have been to calculate the final distribution from the information $x_1 = n_1$,

$$f(p \,|\, x_1 = n_1, \mathcal{B}) = (n_1 + 1)\, p^{n_1}\,, \tag{5.49}$$

and feed it as initial distribution to the next inference:

$$\begin{aligned} f(p \,|\, x_1 = n_1, x_2 = n_2, \mathcal{B}) &= \frac{p^{n_2} f(p \,|\, x_1 = n_1, \mathcal{B})}{\int_0^1 p^{n_2} f(p \,|\, x_1 = n_1, \mathcal{B})\, \mathrm{d}p} && (5.50)\\ &= \frac{p^{n_2} (n_1+1)\, p^{n_1}}{\int_0^1 p^{n_2} (n_1+1)\, p^{n_1}\, \mathrm{d}p} && (5.51)\\ &= (n_1 + n_2 + 1)\, p^{n_1+n_2}\,, && (5.52) \end{aligned}$$

getting the same result.

5.5.2 Poisson distributed quantities

As is well known, the typical application of the Poisson distribution is in counting experiments such as source activity, cross-sections, etc. The unknown parameter to be inferred is λ. Applying the Bayes formula we get

$$f(\lambda \,|\, x, \mathcal{P}) = \frac{\frac{\lambda^x\, e^{-\lambda}}{x!}\, f_\circ(\lambda)}{\int_0^\infty \frac{\lambda^x\, e^{-\lambda}}{x!}\, f_\circ(\lambda)\, \mathrm{d}\lambda}\,. \tag{5.53}$$

Assuming[7] $f_\circ(\lambda)$ constant up to a certain $\lambda_{max} \gg x$ and making the integral by parts we obtain

$$\begin{aligned} f(\lambda \,|\, x, \mathcal{P}) &= \frac{\lambda^x\, e^{-\lambda}}{x!}\,, && (5.54)\\ F(\lambda \,|\, x, \mathcal{P}) &= 1 - e^{-\lambda} \left(\sum_{n=0}^{x} \frac{\lambda^n}{n!} \right)\,, && (5.55) \end{aligned}$$

where the last result has been obtained by integrating (5.54) also by parts. Figure 5.3 shows how to build the credibility intervals, given a certain measured number of counts x. Figure 5.4 shows some numerical examples.

$f(\lambda)$ has the following properties.

- Expected value, variance, and value of maximum probability are

$$\begin{aligned} \mathrm{E}[\lambda] &= x + 1, && (5.56)\\ \mathrm{Var}(\lambda) &= x + 1, && (5.57)\\ \lambda_m &= x\,. && (5.58) \end{aligned}$$

[7] There is a school of thought according to which the most appropriate function is $f_\circ(\lambda) \propto 1/\lambda$. If you think that it is reasonable for your problem, it may be a good prior. Claiming that this is 'the truth' is one of the many claims of the angels' sex determinations. For didactical purposes a uniform distribution is more than enough. Some comments about the $1/\lambda$ prescription will be given when discussing the particular case $x = 0$.
Note added: criticisms concerning so-called 'reference priors' can be found in Ref.[22].

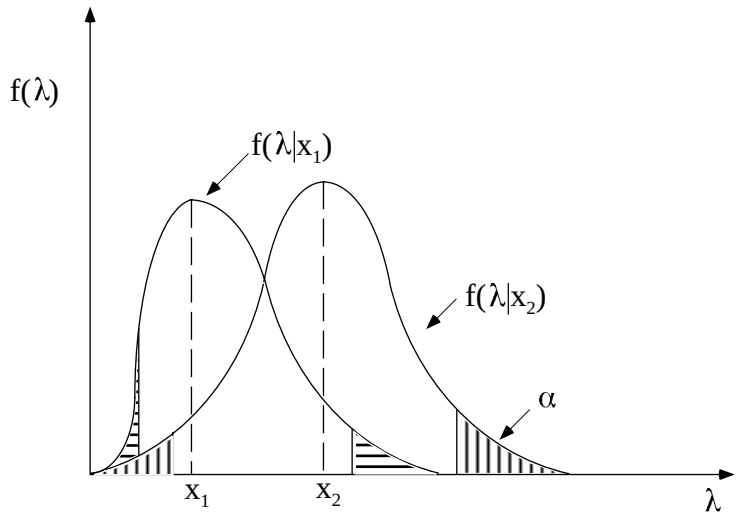

Figure 5.3: Poisson parameter λ inferred from an observed number x of counts.

The fact that the best estimate of λ in the Bayesian sense is not the intuitive value x but $x+1$ should neither surprise, nor disappoint us. According to the initial distribution used there are always more possible values of λ on the right side than on the left side of x, and they pull the distribution to their side; the full information is always given by $f(\lambda)$ and the use of the mean is just a rough approximation; the difference from the desired intuitive value x in units of the standard deviation goes as $1/\sqrt{x+1}$ and becomes immediately negligible.

- When x becomes large we get

$$\begin{aligned} \mathrm{E}[\lambda] &\approx \lambda_m = x\,, & (5.59)\\ \mathrm{Var}(\lambda) &\approx \lambda_m = x\,, & (5.60)\\ \sigma_\lambda &\approx \sqrt{x}\,, & (5.61)\\ \lambda &\sim \mathcal{N}(x, \sqrt{x})\,. & (5.62) \end{aligned}$$

Equation (5.61) is one of the most familar formulae used by physicists to assess the uncertainty of a measurement, although it is sometimes misused.

Let us conclude with a special case: $x = 0$ (see Fig. 5.5). As one might imagine, the inference is highly sensitive to the initial distribution. Let us assume that the experiment was planned with the hope of observing something, i.e. that it could detect a handful of events within its lifetime. With this hypothesis one may use any vague prior function not strongly peaked at zero. We have already come across a similar case in Section 5.4.3, concerning the upper limit of the neutrino mass. There it was shown that reasonable hypotheses based on the positive attitude of the experimentalist are almost equivalent and that they give results consistent with detector performances. Let us use then the uniform distribution

$$\begin{aligned} f(\lambda \,|\, x=0, \mathcal{P}) &= e^{-\lambda}, & (5.63)\\ F(\lambda \,|\, x=0, \mathcal{P}) &= 1 - e^{-\lambda}, & (5.64)\\ \lambda &< 3 \text{ at } 95\% \text{ probability}\,. & (5.65) \end{aligned}$$

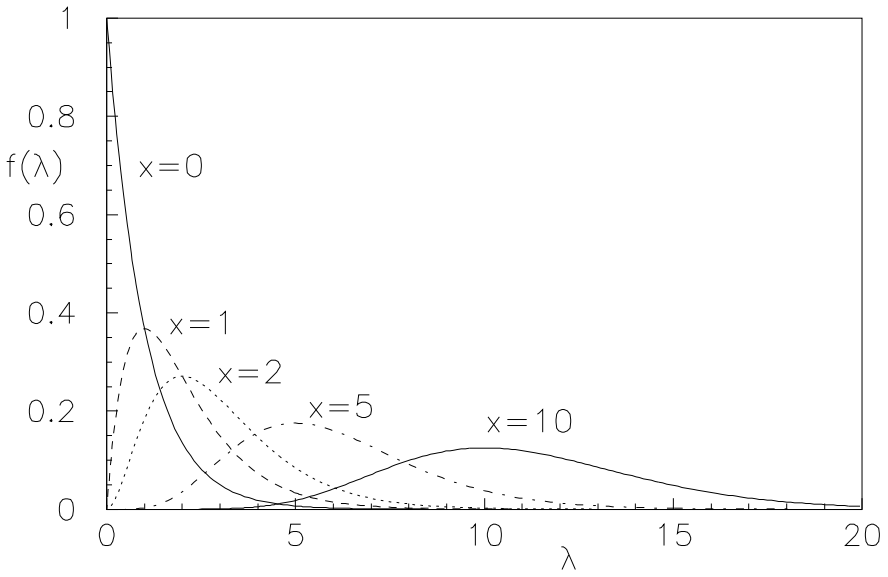

Figure 5.4: Examples of $f(\lambda \mid x_i)$.

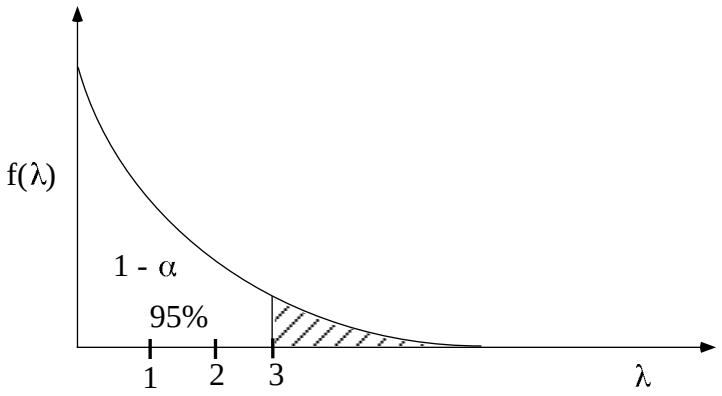

Figure 5.5: Upper limit to λ having observed 0 events.

5.6 Uncertainty due to systematic errors of unknown size

5.6.1 Example: uncertainty of the instrument scale offset

In our scheme any quantity of influence of which we do not know the exact value is a source of systematic error. It will change the final distribution of μ and hence its uncertainty. We have already discussed the most general case in Section 5.2.1. Let us make a simple application making a small variation to the example in Section 5.4.1: the 'zero' of the instrument is not known exactly, owing to calibration uncertainty. This can be parametrized assuming that its true value Z is normally distributed around 0 (i.e. the calibration was properly done!) with a standard deviation σ_Z. Since, most probably, the true value of μ is independent of the true value of Z, the initial joint probability density function can be written as the product of the

marginal ones:

$$f_\circ(\mu, z) = f_\circ(\mu)\, f_\circ(z) = k\, \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right]. \tag{5.66}$$

Also the likelihood changes with respect to (5.11):

$$f(x_1 \,|\, \mu, z) = \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1 - \mu - z)^2}{2\,\sigma_1^2}\right]. \tag{5.67}$$

Putting all the pieces together and making use of (5.3) we finally get

$$f(\mu \,|\, x_1, \ldots, f_\circ(z)) = \frac{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}z}{\iint \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}\mu\, \mathrm{d}z}.$$

Integrating[8] we get

$$f(\mu) = f(\mu \,|\, x_1, \ldots, f_\circ(z)) = \frac{1}{\sqrt{2\,\pi}\,\sqrt{\sigma_1^2 + \sigma_Z^2}}\, \exp\left[-\frac{(\mu - x_1)^2}{2\,(\sigma_1^2 + \sigma_Z^2)}\right]. \tag{5.68}$$

The result is that $f(\mu)$ is still a Gaussian, but with a larger variance. The global standard uncertainty is the quadratic combination of that due to the statistical fluctuation of the data sample and the uncertainty due to the imperfect knowledge of the systematic effect:

$$\sigma_{tot}^2 = \sigma_1^2 + \sigma_Z^2\,. \tag{5.69}$$

This result is well known, although there are still some old-fashioned recipes which require different combinations of the contributions to be performed.

It must be noted that in this framework it makes no sense to speak of statistical and systematic uncertainties, as if they were of a different nature. They have the same probabilistic nature: $\overline{Q}_{n_1}$ is around μ with a standard deviation σ_1, and Z is around 0 with standard deviation σ_Z. What distinguishes the two components is how the knowledge of the uncertainty is gained: in one case (σ_1) from repeated measurements on the physics quantity of interest; in the second case (σ_Z) the evaluation was done by somebody else (the constructor of the instrument), or in a previous experiment, or guessed from the knowledge of the detector, or by simulation, etc. This is the reason why the ISO Guide [3] prefers the generic names 'type A' and 'type B' for the two kinds of contribution to global uncertainty. In particular, the name 'systematic uncertainty' should be avoided, while it is correct to speak about 'uncertainty due to a systematic effect'.

5.6.2 Correction for known systematic errors

It is easy to be convinced that if our prior knowledge about Z was of the kind

$$Z \sim \mathcal{N}(z_\circ, \sigma_Z) \tag{5.70}$$

[8] It may help to know that

$$\int_{-\infty}^{+\infty} \exp\left[b\,x - \frac{x^2}{a^2}\right] \mathrm{d}x = \sqrt{a^2\,\pi}\, \exp\left[\frac{a^2\,b^2}{4}\right].$$

the result would have been

$$\mu \sim \mathcal{N}\left(x_1 - z_\circ, \sqrt{\sigma_1^2 + \sigma_Z^2}\right) , \tag{5.71}$$

i.e. one has first to correct the result for the best value of the systematic error and then include in the global uncertainty a term due to imperfect knowledge about it. This is a well-known and practised procedure, although there are still people who confuse $z_\circ$ with its uncertainty.

5.6.3 Measuring two quantities with the same instrument having an uncertainty of the scale offset

Let us take an example which is a little more complicated (at least from the mathematical point of view) but conceptually very simple and also very common in laboratory practice. We measure two physical quantities with the same instrument, assumed to have an uncertainty on the 'zero', modelled with a normal distribution as in the previous sections. For each of the quantities we collect a sample of data under the same conditions, which means that the unknown offset error does not change from one set of measurements to the other. Calling μ_1 and μ_2 the true values, x_1 and x_2 the sample averages, σ_1 and σ_2 the average's standard deviations, and Z the true value of the zero, the initial probability density and the likelihood are

$$f_\circ(\mu_1, \mu_2, z) = f_\circ(\mu_1)\, f_\circ(\mu_2)\, f_\circ(z) = k\, \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \tag{5.72}$$

and

$$\begin{aligned} f(x_1, x_2 \,|\, \mu_1, \mu_2, z) &= \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1 - \mu_1 - z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_2}\, \exp\left[-\frac{(x_2 - \mu_2 - z)^2}{2\,\sigma_2^2}\right] \\ &= \frac{1}{2\,\pi\,\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{(x_1 - \mu_1 - z)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2 - z)^2}{\sigma_2^2}\right)\right], \end{aligned} \tag{5.73}$$

respectively. The result of the inference is now the joint probability density function of μ_1 and μ_2:

$$f(\mu_1, \mu_2 \,|\, x_1, x_2, \sigma_1, \sigma_2, f_\circ(z)) = \frac{\int f(x_1, x_2 \,|\, \mu_1, \mu_2, z)\, f_\circ(\mu_1, \mu_2, z) \mathrm{d}z}{\iiint f(x_1, x_2 \,|\, \mu_1, \mu_2, z)\, f_\circ(\mu_1, \mu_2, z)\, \mathrm{d}\mu_1\, \mathrm{d}\mu_2\, \mathrm{d}z} , \tag{5.74}$$

where expansion of the functions has been omitted for the sake of clarity. Integrating we get

$$\begin{aligned} f(\mu_1, \mu_2) &= \frac{1}{2\,\pi\,\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}\,\sqrt{1 - \rho^2}} \\ &\quad \exp\left\{-\frac{1}{2\,(1 - \rho^2)}\left[\frac{(\mu_1 - x_1)^2}{\sigma_1^2 + \sigma_Z^2} - 2\,\rho\,\frac{(\mu_1 - x_1)(\mu_2 - x_2)}{\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}} + \frac{(\mu_2 - x_2)^2}{\sigma_2^2 + \sigma_Z^2}\right]\right\}. \end{aligned} \tag{5.75}$$

where

$$\rho = \frac{\sigma_Z^2}{\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}} . \tag{5.76}$$

If σ_Z vanishes then (5.75) has the simpler expression

$$f(\mu_1,\mu_2) \xrightarrow[\sigma_Z\to 0]{} [\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\,\exp\left[-\frac{(\mu_1-x_1)^2}{2\,\sigma_1^2}\right]\frac{1}{\sqrt{2\,\pi}\,\sigma_2}\,\exp\left[-\frac{(\mu_2-x_2)^2}{2\,\sigma_2^2}\right]]$$

i.e. if there is no uncertainty on the offset calibration then the joint density function $f(\mu_1,\mu_2)$ is equal to the product of two independent normal functions, i.e. μ_1 and μ_2 are independent. In the general case we have to conclude the following.

- The effect of the common uncertainty σ_Z makes the two values correlated, since they are affected by a common unknown systematic error; the correlation coefficient is always non-negative ($\rho \geq 0$), as intuitively expected from the definition of systematic error.

- The joint density function is a multinormal distribution of parameters x_1, $\sigma_{\mu_1} = \sqrt{\sigma_1^2+\sigma_Z^2}$, x_2, $\sigma_{\mu_2} = \sqrt{\sigma_2^2+\sigma_Z^2}$, and ρ (see example of Fig. 4.1).

- The marginal distributions are still normal:

$$\mu_1 \sim \mathcal{N}\left(x_1, \sqrt{\sigma_1^2+\sigma_Z^2}\right), \tag{5.77}$$

$$\mu_2 \sim \mathcal{N}\left(x_2, \sqrt{\sigma_2^2+\sigma_Z^2}\right). \tag{5.78}$$

- The covariance between μ_1 and μ_2 is

$$\begin{aligned}\mathrm{Cov}(\mu_1,\mu_2) &= \rho\,\sigma_{\mu_1}\sigma_{\mu_2} \\ &= \rho\sqrt{\sigma_1^2+\sigma_Z^2}\sqrt{\sigma_2^2+\sigma_Z^2} = \sigma_Z^2\,.\end{aligned} \tag{5.79}$$

- The distribution of any function $g(\mu_1,\mu_2)$ can be calculated using the standard methods of probability theory. For example, one can demonstrate that the sum $S=\mu_1+\mu_2$ and the difference $D=\mu_1-\mu_2$ are also normally distributed (see also the introductory discussion to the central limit theorem and Section 6.3 for the calculation of averages and standard deviations):

$$S \sim \mathcal{N}\left(x_1+x_2, \sqrt{\sigma_1^2+\sigma_2^2+(2\,\sigma_Z)^2}\right), \tag{5.80}$$

$$D \sim \mathcal{N}\left(x_1-x_2, \sqrt{\sigma_1^2+\sigma_2^2}\right). \tag{5.81}$$

  The result can be interpreted in the following way.

  - The uncertainty on the difference does not depend on the common offset uncertainty: whatever the value of the true zero is, it cancels in differences.
  - In the sum, instead, the effect of the common uncertainty is somewhat amplified since it enters 'in phase' in the global uncertainty of each of the quantities.

5.6.4 Indirect calibration

Let us use the result of the previous section to solve another typical problem of measurements. Suppose that after (or before, it doesn't matter) we have done the measurements of x_1 and x_2 and we have the final result, summarized in (5.75), we know the exact value of μ_1 (for example we perform the measurement on a reference). Let us call it μ_1°. Will this information provide a better knowledge of μ_2? In principle yes: the difference between x_1 and μ_1° defines the systematic error (the true value of the zero Z). This error can then be subtracted from x_2 to get a corrected value. Also the overall uncertainty of μ_2 should change, intuitively it should decrease, since we are adding new information. But its value doesn't seem to be obvious, since the logical link between μ_1° and μ_2 is $\mu_1^\circ \to Z \to \mu_2$.

The problem can be solved exactly using the concept of conditional probability density function $f(\mu_2 \,|\, \mu_1^\circ)$ [see (4.70)–(4.71)]. We get

$$\mu_{2\,|\,\mu_1^\circ} \sim \mathcal{N}\left(x_2 + \frac{\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}\,(\mu_1^\circ - x_1),\ \sqrt{\sigma_2^2 + \left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_Z^2}\right)^{-1}} \right) . \tag{5.82}$$

The best value of μ_2 is shifted by an amount Δ, with respect to the measured value x_2, which is not exactly $x_1 - \mu_1^\circ$, as was naïvely guessed, and the uncertainty depends on σ_2, σ_Z and σ_1. It is easy to be convinced that the exact result is more reasonable than the (suggested) first guess. Let us rewrite Δ in two different ways:

$$\Delta = \frac{\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}\,(\mu_1^\circ - x_1) \tag{5.83}$$

$$= \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_Z^2}} \left[\frac{1}{\sigma_1^2}\cdot(x_1-\mu_1^\circ) + \frac{1}{\sigma_Z^2}\cdot 0 \right] . \tag{5.84}$$

- Equation (5.83) shows that one has to apply the correction $x_1 - \mu_1^\circ$ only if $\sigma_1 = 0$. If instead $\sigma_Z = 0$ there is no correction to be applied, since the instrument is perfectly calibrated. If $\sigma_1 \approx \sigma_Z$ the correction is half of the measured difference between x_1 and μ_1°.

- Equation (5.84) shows explicitly what is going on and why the result is consistent with the way we have modelled the uncertainties. In fact we have performed two independent calibrations: one of the offset and one of μ_1. The best estimate of the true value of the zero Z is the weighted average of the two measured offsets.

- The new uncertainty of μ_2 [see (5.82)] is a combination of σ_2 and the uncertainty of the weighted average of the two offsets. Its value is smaller than it would be with only one calibration and, obviously, larger than that due to the sampling fluctuations alone:

$$\sigma_2 \le \sqrt{\sigma_2^2 + \frac{\sigma_1^2\,\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}} \le \sqrt{\sigma_2^2+\sigma_Z^2}\,. \tag{5.85}$$

5.6.5 Counting measurements in the presence of background

As an example of a different kind of systematic effect, let us think of counting experiments in the presence of background. For example we are searching for a new particle, we make some selection cuts and count x events. But we also expect an average number of background events $\lambda_{B_\circ} \pm \sigma_B$, where σ_B is the standard uncertainty of $\lambda_{B_\circ}$, not to be confused with $\sqrt{\lambda_{B_\circ}}$. What can we say about λ_S, the true value of the average number associated with the signal? First we

will treat the case in which the determination of the expected number of background events is well known ($\sigma_B/\lambda_{B_\circ} \ll 1$), and then the general case.

$\sigma_B/\lambda_{B_\circ} \ll 1$: The true value of the sum of signal and background is $\lambda = \lambda_S + \lambda_{B_\circ}$. The likelihood is

$$P(x \,|\, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\,. \tag{5.86}$$

Applying Bayes' theorem we have

$$f(\lambda_S \,|\, x, \lambda_{B_\circ}) \;=\; \frac{e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)}{\int_0^\infty e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)\,\mathrm{d}\lambda_S}\,. \tag{5.87}$$

Choosing again $f_\circ(\lambda_S)$ uniform (in a reasonable interval) this gets simplified. The integral in the denominator can be calculated easily by parts and the final result is

$$f(\lambda_S \,|\, x, \lambda_{B_\circ}) \;=\; \frac{e^{-\lambda_S}\,(\lambda_{B_\circ}+\lambda_S)^x}{x!\,\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}, \tag{5.88}$$

$$F(\lambda_S \,|\, x, \lambda_{B_\circ}) \;=\; 1 - \frac{e^{-\lambda_S}\,\sum_{n=0}^{x}\frac{(\lambda_{B_\circ}+\lambda_S)^n}{n!}}{\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}\,. \tag{5.89}$$

From (5.88) and (5.89) it is possible to calculate in the usual way the best estimate and the credibility intervals of λ_S. Two particular cases are of interest:

- If $\lambda_{B_\circ} = 0$ then formulae (5.54) and (5.55) are recovered. In such a case one measured count is enough to claim for a signal (if somebody is willing to believe that really $\lambda_{B_\circ} = 0$ without any uncertainty ...).
- If $x = 0$ then

  $$f(\lambda \,|\, x, \lambda_{B_\circ}) = e^{-\lambda_S}\,, \tag{5.90}$$

  independently of $\lambda_{B_\circ}$. This result is not really obvious.

Any $g(\lambda_{B_\circ})$: In the general case, the true value of the average number of background events λ_B is unknown. We only known that it is distributed around $\lambda_{B_\circ}$ with standard deviation σ_B and probability density function $g(\lambda_B)$, not necessarily a Gaussian. What changes with respect to the previous case is the initial distribution, now a joint function of λ_S and of λ_B. Assuming λ_B and λ_S independent the prior density function is

$$f_\circ(\lambda_S, \lambda_B) = f_\circ(\lambda_S)\, g_\circ(\lambda_B)\,. \tag{5.91}$$

We leave $f_\circ$ in the form of a joint distribution to indicate that the result we shall get is the most general for this kind of problem. The likelihood, on the other hand, remains the same as in the previous example. The inference of λ_S is done in the usual way, applying Bayes' theorem and marginalizing with respect to λ_S:

$$f(\lambda_S \,|\, x, g_\circ(\lambda_B)) = \frac{\int e^{-(\lambda_B+\lambda_S)}\,(\lambda_B+\lambda_S)^x\, f_\circ(\lambda_S, \lambda_B)\,\mathrm{d}\lambda_\mathrm{B}}{\iint e^{-(\lambda_B+\lambda_S)}\,(\lambda_B+\lambda_S)^x\, f_\circ(\lambda_S, \lambda_B)\,\mathrm{d}\lambda_S\,\mathrm{d}\lambda_B}\,. \tag{5.92}$$

The previous case [formula (5.88)] is recovered if the only value allowed for λ_B is $\lambda_{B_\circ}$ and $f_\circ(\lambda_S)$ is uniform:

$$f_\circ(\lambda_S, \lambda_B) = k\,\delta(\lambda_B - \lambda_{B_\circ})\,. \tag{5.93}$$

Chapter 6

Bypassing Bayes' theorem for routine applications

"Let us consider a dimensionless mass,
suspended from an inextensible massless wire,
free to oscillate without friction ... "
(Any textbook)

6.1 Approximate methods

6.1.1 Linearization

We have seen in the above examples how to use the general formula (5.3) for practical applications. Unfortunately, when the problem becomes more complicated one starts facing integration problems. For this reason approximate methods are generally used. We will derive the approximation rules consistently with the approach followed in these notes and then the resulting formulae will be compared with the ISO recommendations. To do this, let us neglect for a while all quantities of influence which could produce unknown systematic errors. In this case (5.3) can be replaced by (5.4), which can be further simplified if we remember that correlations between the results are originated by unknown systematic errors. In the absence of these, the joint distribution of all quantities $\underline{\mu}$ is simply the product of marginal ones:

$$f_{R_i}(\underline{\mu_i}) = \prod_i f_{R_i}(\mu_i)\,, \tag{6.1}$$

with

$$f_{R_i}(\mu_i) = f_{R_i}(\mu_i \,|\, x_i, \underline{h}_\circ) = \frac{f(x_i \,|\, \mu_i, \underline{h}_\circ)\, f_\circ(\mu_i)}{\int f(x_i \,|\, \mu_i, \underline{h}_\circ)\, f_\circ(\mu_i)\, \mathrm{d}\mu_\mathrm{i}}\,. \tag{6.2}$$

The symbol $f_{R_i}(\mu_i)$ indicates that we are dealing with raw values[1] evaluated at $\underline{h} = \underline{h}_\circ$. Since for any variation of $\underline{h}$ the inferred values of μ_i will change, it is convenient to name with the same subscript R the quantity obtained for $\underline{h}_\circ$:

$$f_{R_i}(\mu_i) \longrightarrow f_{R_i}(\mu_{R_i})\,. \tag{6.3}$$

[1]The choice of the adjective 'raw' will become clearer later on. The subscript R is also meant to represent 'random', in the sense that only sampling effects are considered at the moment.

Let us indicate with $\widehat{\mu}_{R_i}$ and σ_{R_i} the best estimates and the standard uncertainty of the raw values:

$$\widehat{\mu}_{R_i} = \mathrm{E}[\mu_{R_i}]\,, \tag{6.4}$$
$$\sigma^2_{R_i} = \mathrm{Var}(\mu_{R_i})\,. \tag{6.5}$$

For any possible configuration of conditioning hypotheses $\underline{h}$, corrected values μ_i are obtained:

$$\mu_i = \mu_{R_i} + g_i(\underline{h})\,. \tag{6.6}$$

The function which relates the corrected value to the raw value and to the systematic effects has been denoted by g_i so as not to be confused with a probability density function. Expanding (6.6) in series around $\underline{h}_\circ$ we finally arrive at the expression which will allow us to make the approximated evaluations of uncertainties:

$$\boxed{\mu_i = \mu_{R_i} + \sum_l \frac{\partial g_i}{\partial h_l}\,(h_l - h_{\circ_l}) + \ldots} \tag{6.7}$$

(All derivatives are evaluated at $\{\widehat{\mu}_{R_i}, \underline{h}_\circ\}$. To simplify the notation a similar convention will be used in the following formulae.)

Neglecting the terms of the expansion above the first order, and taking the expected values, we get

$$\begin{aligned}
\widehat{\mu}_i &= \mathrm{E}[\mu_i] \\
&\approx \widehat{\mu}_{R_i}\,; && (6.8)\\
\sigma^2_{\mu_i} &= \mathrm{E}\left[(\mu_i - \mathrm{E}[\mu_i])^2\right] \\
&\approx \sigma^2_{R_i} + \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)^2 \sigma^2_{h_l} \\
&\quad \left\{ + 2 \sum_{l<m} \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_i}{\partial h_m}\right) \rho_{lm}\,\sigma_{h_l}\,\sigma_{h_m} \right\}\,; && (6.9)\\
\mathrm{Cov}(\mu_i, \mu_j) &= \mathrm{E}\left[\,(\mu_i - \mathrm{E}[\mu_i])(\mu_j - \mathrm{E}[\mu_j])\,\right] \\
&\approx \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_l}\right) \sigma^2_{h_l} \\
&\quad \left\{ + 2 \sum_{l<m} \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_m}\right) \rho_{lm}\,\sigma_{h_l}\,\sigma_{h_m} \right\}\,. && (6.10)
\end{aligned}$$

The terms included within $\{\cdot\}$ vanish if the unknown systematic errors are uncorrelated, and the formulae become simpler. Unfortunately, very often this is not the case, as when several calibration constants are simultaneously obtained from a fit (for example, in most linear fits slope and intercept have a correlation coefficient close to -0.9).

Sometimes the expansion (6.7) is not performed around the best values of $\underline{h}$ but around their nominal values, in the sense that the correction for the known value of the systematic errors has not yet been applied (see Section 5.6.2). In this case (6.7) should be replaced by

$$\mu_i = \mu_{R_i} + \sum_l \frac{\partial g_i}{\partial h_l}\,(h_l - h_{N_l}) + \ldots\,, \tag{6.11}$$

where the subscript N stands for nominal. The best value of μ_i is then

$$\begin{aligned}\widehat{\mu}_i &= \mathrm{E}[\mu_i] \\ &\approx \widehat{\mu}_{R_i} + \mathrm{E}\left[\sum_l \frac{\partial g_i}{\partial h_l}\,(h_l - h_{N_l})\right] \\ &= \widehat{\mu}_{R_i} + \sum_l \delta\mu_{i_l}\,. \end{aligned} \tag{6.12}$$

(6.9) and (6.10) instead remain valid, with the condition that the derivative is calculated at $\underline{h}_N$. If $\rho_{lm} = 0$, it is possible to rewrite (6.9) and (6.10) in the following way, which is very convenient for practical applications:

$$\sigma^2_{\mu_i} \approx \sigma^2_{R_i} + \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)^2 \sigma^2_{h_l} \tag{6.13}$$

$$= \sigma^2_{R_i} + \sum_l u^2_{i_l}\,; \tag{6.14}$$

$$\mathrm{Cov}(\mu_i, \mu_j) \approx \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_l}\right)\sigma^2_{h_l} \tag{6.15}$$

$$= \sum_l s_{ij_l} \left|\frac{\partial g_i}{\partial h_l}\right| \sigma_{h_l} \left|\frac{\partial g_j}{\partial h_l}\right| \sigma_{h_l} \tag{6.16}$$

$$= \sum_l s_{ij_l}\, u_{i_l}\, u_{j_l} \tag{6.17}$$

$$= \sum_l \mathrm{Cov}_l(\mu_i, \mu_j)\,. \tag{6.18}$$

u_{i_l} is the component of the standard uncertainty due to effect h_l. s_{ij_l} is equal to the product of signs of the derivatives, which takes into account whether the uncertainties are positively or negatively correlated.

To summarize, when systematic effects are not correlated with each other, the following quantities are needed to evaluate the corrected result, the combined uncertainties and the correlations:

- the raw $\widehat{\mu}_{R_i}$ and σ_{R_i};
- the best estimates of the corrections $\delta\mu_{i_l}$ for each systematic effect h_l;
- the best estimate of the standard deviation u_{i_l} due to the imperfect knowledge of the systematic effect;
- for any pair $\{\mu_i, \mu_j\}$ the sign of the correlation s_{ij_l} due to the effect h_l.

In HEP applications it is frequently the case that the derivatives appearing in (6.12)–(6.16) cannot be calculated directly, as for example when h_l are parameters of a simulation program, or acceptance cuts. Then variations of $\underline{\mu}_i$ are usually studied by varying a particular h_l within a reasonable interval, holding the other influence quantities at the nominal value. $\delta\mu_{i_l}$ and u_{i_l} are calculated from the interval $\pm\Delta^{\pm}_i$ of variation of the true value for a given variation $\pm\Delta^{\pm}_{h_l}$ of h_l and from the probabilistic meaning of the intervals (i.e. from the assumed distribution of the true value). This empirical procedure for determining $\delta\mu_{i_l}$ and u_{i_l} has the advantage that

it can take into account nonlinear effects [45], since it directly measures the difference $\widehat{\mu}_i - \widehat{\mu}_{R_i}$ for a given difference $h_l - h_{N_l}$.

Some examples are given in Section 6.1.4, and two typical experimental applications will be discussed in more detail in Section 6.3.

6.1.2 BIPM and ISO recommendations

In this section we compare the results obtained in the previous section with the recommendations of the Bureau International des Poids et Mesures (BIPM) and the International Organization for Standardization (ISO) on the expression of experimental uncertainty (Refs. [2, 3]).

1. *"The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:*

   A: *those which are evaluated by statistical methods;*

   B: *those which are evaluated by other means.*

   There is not always a simple correspondence between the classification into categories A or B and the previously used classification into 'random' and 'systematic' uncertainties. The term 'systematic uncertainty' can be misleading and should be avoided.

   The detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical result."

   Essentially the first recommendation states that all uncertainties can be treated probabilistically. The distinction between types A and B is subtle and can be misleading if one thinks of statistical methods as synonymous with probabilistic methods, as is currently the case in HEP. Here 'statistical' has the classical meaning of repeated measurements.

2. *"The components in category A are characterized by the estimated variances s_i^2 (or the estimated "standard deviations" s_i) and the number of degrees of freedom ν_i. Where appropriate, the covariances should be given."*

   The estimated variances correspond to $\sigma^2_{R_i}$ of the previous section. The degrees of freedom are related to small samples and to the Student t-distribution. The problem of small samples is not discussed in these notes, but clearly this recommendation is a relic of frequentistic methods. With the approach followed in these notes there is no need to talk about degrees of freedom, since the Bayesian inference defines the final probability function $f(\mu)$ completely.[2]

3. *"The components in category B should be characterized by quantities u_j^2, which may be considered as approximations to the corresponding variances, the existence of which is assumed. The quantities u_j^2 may be treated like variances and the quantities u_j like standard deviations. Where appropriate, the covariances should be treated in a similar way."*

   Clearly, this recommendation is meaningful only in a Bayesian framework.

4. *"The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of 'standard deviations'."*

   This is what we have found in (6.9) and (6.10).

[2] **Note added**: for criticisms about the standard treatment of the small-sample problem see Ref. [22].

5. *"If, for particular applications, it is necessary to multiply the combined uncertainty by a factor to obtain an overall uncertainty, the multiplying factor used must always be stated."*

   This last recommendation states once more that the uncertainty is by default the standard deviation of the true value distribution. Any other quantity calculated to obtain a credibility interval with a certain probability level should be clearly stated.

To summarize, the following are the basic ingredients of the BIPM/ISO recommendations.

subjective definition of probability: it allows variances to be assigned conceptually to any physical quantity which has an uncertain value;

uncertainty as standard deviation:

- it is standard;
- the rule of combination (4.62)–(4.66) applies to standard deviations and not to confidence intervals;

combined standard uncertainty: it is obtained by the usual formula of error propagation and it makes use of variances, covariances and first derivatives;

central limit theorem: it makes, under proper conditions, the true value normally distributed if one has several sources of uncertainty.

Consultation of the ISO Guide [3] is recommended for further explanations about the justification of the standards, for the description of evaluation procedures, and for examples. I would just like to end this section with some examples of the evaluation of type B uncertainties and with some words of caution concerning the use of approximations and of linearization.

6.1.3 Evaluation of type B uncertainties

The ISO Guide states that

"For estimate x_i of an input quantity[3] X_i that has not been obtained from repeated observations, the ... standard uncertainty u_i is evaluated by scientific judgement based on all the available information on the possible variability of X_i. The pool of information may include

- *previous measurement data;*
- *experience with or general knowledge of the behaviour and properties of relevant materials and instruments;*
- *manufacturer's specifications;*
- *data provided in calibration and other certificates;*
- *uncertainties assigned to reference data taken from handbooks."*

6.1.4 Examples of type B uncertainties

1. Previous measurements of other particular quantities, performed in similar conditions, have provided a repeatability standard deviation[4] of σ_r:

$$u = \sigma_r \, .$$

[3] By 'input quantity' the ISO Guide means any of the contributions h_l or μ_{R_i} which enter into (6.9) and (6.10).
[4] This example shows a type B uncertainty originated by random errors.

2. A manufacturer's calibration certificate states that the uncertainty, defined as k standard deviations, is $\pm\Delta$:
$$u = \frac{\Delta}{k}\,.$$

3. A result is reported in a publication as $\overline{x} \pm \Delta$, stating that the average has been performed on four measurements and the uncertainty is a 95% confidence interval. One has to conclude that the confidence interval has been calculated using the Student t-distribution:
$$u = \frac{\Delta}{3.18}\,.$$

4. A manufacturer's specification states that the error on a quantity should not exceed Δ. With this limited information one has to assume a uniform distribution:
$$u = \frac{2\,\Delta}{\sqrt{12}} = \frac{\Delta}{\sqrt{3}}\,.$$

5. A physical parameter of a Monte Carlo is believed to lie in the interval of $\pm\Delta$ around its best value, but not with uniform distribution: the degree of belief that the parameter is at the centre is higher than the degree of belief that it is at the edges of the interval. With this information a triangular distribution can be reasonably assumed:
$$u = \frac{\Delta}{\sqrt{6}}\,.$$
Note that the coefficient in front of Δ changes from the 0.58 of the previous example to the 0.41 of this. If the interval $\pm\Delta$ were a 3σ interval then the coefficient would have been equal to 0.33. These variations — to be considered extreme — are smaller than the statistical fluctuations of empirical standard deviations estimated from ≈ 10 measurements. This shows that one should not be worried that the type B uncertainties are less accurate than type A, especially if one tries to model the distribution of the physical quantity honestly.

6. The absolute energy calibration of an electromagnetic calorimeter module is not known exactly and is estimated to be between the nominal one and +10%. The statistical error is known by test beam measurements to be $18\%/\sqrt{E/\mathrm{GeV}}$. What is the uncertainty on the energy measurement of an electron which has apparently released 30 GeV?

   - There is no type A uncertainty, since only one measurement has been performed.
   - The energy has to be corrected for the best estimate of the calibration constant: +5%, with a relative uncertainty of $18\%/\sqrt{31.5}$ due to sampling (the statistical error):
   $$E = 31.5 \pm 1.0\,\mathrm{GeV}\,.$$
   - Then one has to take into account the uncertainty due to absolute energy scale calibration:
     - assuming a uniform distribution of the true calibration constant, $u = 31.5 \times 0.1/\sqrt{12} = 0.9\,\mathrm{GeV}$:
     $$E = 31.5 \pm 1.3\,\mathrm{GeV}\,;$$
     - assuming, more reasonably, a triangular distribution, $u = 31.5 \times 0.05/\sqrt{6} = 0.6\,\mathrm{GeV}$,
     $$E = 31.5 \pm 1.2\,\mathrm{GeV}\,.$$

- Instead, interpreting the maximum deviation from the nominal calibration as uncertainty (see comment at the end of Section 5.6.2),

$$E = 30.0 \pm 1.0 \pm 3.0\,\mathrm{GeV} \rightarrow E = 30.0 \pm 3.2\,\mathrm{GeV}\,.$$

  As already mentioned earlier in these notes, while reasonable assumptions (in this case the first two) give consistent results, this is not true if one makes inconsistent use of the information just for the sake of giving safe uncertainties.

7. **Note added**: the original version of the primer contained at this point a more realistic and slightly more complicated example, which requires, instead, a next-to-linear treatment [45], which was not included in the notes, neither is it in this new version. Therefore, I prefer to skip this example in order to avoid confusion.

6.1.5 Caveat concerning the blind use of approximate methods

The mathematical apparatus of variances and covariances of (6.9) and (6.10) is often seen as the most complete description of uncertainty and in most cases used blindly in further uncertainty calculations. It must be clear, however, that this is just an approximation based on linearization. If the function which relates the corrected value to the raw value and the systematic effects is not linear then the linearization may cause trouble. An interesting case is discussed in Section 6.3.

There is another problem which may arise from the simultaneous use of Bayesian estimators and approximate methods. Let us introduce the problem with an example.

Example 1: 1000 independent measurements of the efficiency of a detector have been performed (or 1000 measurements of branching ratio, if you prefer). Each measurement was carried out on a base of 100 events and each time 10 favourable events were observed (this is obviously strange — though not impossible — but it simplifies the calculations). The result of each measurement will be [see (5.33)–(5.35)]:

$$\widehat{\epsilon}_i = \frac{10+1}{100+2} = 0.1078, \tag{6.19}$$

$$\sigma(\epsilon_i) = \sqrt{\frac{11 \times 91}{103 \times 102^2}} = 0.031\,. \tag{6.20}$$

Combining the 1000 results using the standard weighted average procedure gives

$$\epsilon = 0.1078 \pm 0.0010\,. \tag{6.21}$$

Alternatively, taking the complete set of results to be equivalent to 100 000 trials with 10 000 favourable events, the combined result is

$$\epsilon' = 0.10001 \pm 0.0009 \tag{6.22}$$

(the same as if one had used Bayes' theorem iteratively to infer $f(\epsilon)$ from the the partial 1000 results). The conclusions are in disagreement and the first result is clearly mistaken (the solution will be given after the following example).

The same problem arises in the case of inference of the Poisson distribution parameter λ and, in general, whenever $f(\mu)$ is not symmetrical around $\mathrm{E}[\mu]$.

Example 2: Imagine an experiment running continuously for one year, searching for monopoles and identifying none. The consistency with zero can be stated either quoting $\mathrm{E}[\lambda] = 1$ and $\sigma_\lambda = 1$, or a 95% upper limit $\lambda < 3$. In terms of rate (number of monopoles per day) the result would be either $\mathrm{E}[r] = 2.7 \cdot 10^{-3}$, $\sigma(r) = 2.7 \cdot 10^{-3}$, or an upper limit $r < 8.2 \cdot 10^{-3}$. It is easy to show that, if we take the 365 results for each of the running days and combine them using the standard weighted average, we get $r = 1.00 \pm 0.05$ monopoles per day![5] This absurdity is not caused by the Bayesian method, but by the abuse of standard rules for combining the results (the weighted average formulae (5.19) and (5.20) are derived from the normal distribution hypothesis). Using Bayesian inference would have led to a consistent and reasonable result no matter how the 365 days of running had been subdivided for partial analysis.

This suggests that in some cases it could be preferable to present the result also providing the mode of μ (p_m and λ_m of Sections 5.5.1 and 5.5.2). This way of presenting the results is similar to that suggested by the maximum likelihood approach, with the difference that for $f(\mu)$ one should take the final probability density function and not simply the likelihood. Since it is practically impossible to summarize the outcome of an inference in only two numbers (best value and uncertainty), in case of non-normality of the $f(\mu)$, more information about $f(\mu)$ should be given.

6.2 Indirect measurements

Conceptually this is a very simple task in the Bayesian framework, whereas the frequentistic one requires a lot of gymnastics, going back and forth from the logical level of true values to the logical level of estimators. If one accepts that the true values are just random variables,[6] then, calling Y a function of other quantities X, each having a probability density function $f(x)$, the probability density function of Y $f(y)$ can be calculated with the standard formulae which follow from the rules probability. Note that in the approach presented in these notes uncertainties due to systematic effects are treated in the same way as indirect measurements. It is worth repeating that there is no conceptual distinction between various components of the measurement uncertainty. When approximations are sufficient, formulae (6.9) and (6.10) can be used.

Let us take an example for which the linearization does not give the right result.

Example: The speed of a proton is measured with a time-of-flight system. Find the 68, 95 and 99% probability intervals for the energy, knowing that $\beta = v/c = 0.9971$, and that distance and time have been measured with a 0.2% accuracy.

The relation

$$E = \frac{mc^2}{\sqrt{1-\beta^2}}$$

is strongly nonlinear. The results given by the approximated method and the correct one are shown in the table below.

[5] **Note added:** this is exactly the presumed paradox reported by the 1998 issue of the PDG [46] as an argument against Bayesian statistics (Section 29.6.2, p. 175: *"If Bayesian estimates are averaged, they do not converge to the true value, since they have all been forced to be positive."*)

[6] To make the formalism lighter, let us call both the random variable associated with the quantity and the quantity itself by the same name X_i (instead of μ_{x_i}).

Probability (%)	Linearization E (GeV)	Correct result E (GeV)
68	$6.4 \leq E \leq 18$	$8.8 \leq E \leq 64$
95	$0.7 \leq E \leq 24$	$7.2 \leq E < \infty$
99	$0. \leq E \leq 28$	$6.6 \leq E < \infty$

6.3 Covariance matrix of experimental results

This section, based on Ref. [47], shows once more practical rules to build the covariance matrix associated with experimental data with correlated uncertainty (see also Sections 5.6.3 and 6.1.1), treating explicitly also the case of normalization uncertainty. Then it will be shown that, in this case, the covariance matrix evaluated in this way produces biased χ^2 fits.

6.3.1 Building the covariance matrix of experimental data

In physics applications, it is rarely the case that the covariance between the best estimates of two physical quantities,[7] each given by the arithmetic average of direct measurements ($x_i = \overline{X_i} = \frac{1}{n}\sum_{k=1}^{n} X_{ik}$), can be evaluated from the sample covariance[8] of the two averages:

$$\mathrm{Cov}(x_i, x_j) = \frac{1}{n\,(n-1)} \sum_{k=1}^{n} (X_{ik} - \overline{X}_i)(X_{jk} - \overline{X}_j) \ . \tag{6.23}$$

More frequent is the well-understood case in which the physical quantities are obtained as a result of a χ^2 minimization, and the terms of the inverse of the covariance matrix are related to the curvature of χ^2 at its minimum:

$$\left(V^{-1}\right)_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial X_i \partial X_j} \right|_{x_i, x_j} \ . \tag{6.24}$$

In most cases one determines independent values of physical quantities with the same detector, and the correlation between them originates from the detector calibration uncertainties. Frequentistically, the use of (6.23) in this case would correspond to having a sample of detectors, each of which is used to perform a measurement of all the physical quantities.

A way of building the covariance matrix from the direct measurements is to consider the original measurements and the calibration constants as a common set of independent and uncorrelated measurements, and then to calculate corrected values that take into account the calibration constants. The variance/covariance propagation will automatically provide the full covariance matrix of the set of results. Let us derive it for two cases that occur frequently, and then proceed to the general case.

[7] In this section the symbol X_i will indicate the variable associated to the i-th physical quantity and X_{ik} its k-th direct measurement; x_i the best estimate of its value, obtained by an average over many direct measurements or indirect measurements, σ_i the standard deviation, and y_i the value corrected for the calibration constants. The weighted average of several x_i will be denoted by $\overline{x}$.

[8] **Note added:** The '$n-1$' at the denominator of (6.23) is for the same reason as the '$n-1$' of the sample standard deviation. Although I do not agree with the rationale behind it, this formula can be considered a kind of standard and, anyhow, replacing '$n-1$' by 'n' has no effect in normal applications. As already said, in these notes I will not discuss the small-sample problem; anyone who is interested in my worries concerning default formulae for small samples, as well as Student t-distribution may have a look at Ref. [22].

Offset uncertainty

Let $x_i \pm \sigma_i$ be the $i = 1 \ldots n$ results of independent measurements and $\mathbf{V}_X$ the (diagonal) covariance matrix. Let us assume that they are all affected by the same calibration constant c, having a standard uncertainty σ_c. The corrected results are then $y_i = x_i + c$. We can assume, for simplicity, that the most probable value of c is 0, i.e. the detector is well calibrated. One has to consider the calibration constant as the physical quantity X_{n+1}, whose best estimate is $x_{n+1} = 0$. A term $V_{X_{n+1,n+1}} = \sigma_c^2$ must be added to the covariance matrix.

The covariance matrix of the corrected results is given by the transformation

$$\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^T\,, \tag{6.25}$$

where $M_{ij} = \left.\frac{\partial Y_i}{\partial X_j}\right|_{x_j}$. The elements of $\mathbf{V}_Y$ are given by

$$V_{Y_{kl}} = \sum_{ij} \left.\frac{\partial Y_k}{\partial X_i}\right|_{x_i} \left.\frac{\partial Y_l}{\partial X_j}\right|_{x_j} V_{X_{ij}}\,. \tag{6.26}$$

In this case we get

$$\begin{aligned}
\sigma^2(Y_i) &= \sigma_i^2 + \sigma_c^2, & (6.27)\\
\mathrm{Cov}(Y_i, Y_j) &= \sigma_c^2 \qquad (i \neq j), & (6.28)\\
\rho_{ij} &= \frac{\sigma_c^2}{\sqrt{\sigma_i^2 + \sigma_c^2}\sqrt{\sigma_j^2 + \sigma_c^2}} & (6.29)\\
&= \frac{1}{\sqrt{1 + \left(\frac{\sigma_i}{\sigma_c}\right)^2}\sqrt{1 + \left(\frac{\sigma_j}{\sigma_c}\right)^2}}\,, & (6.30)
\end{aligned}$$

reobtaining the results of Section 5.6.3. The total uncertainty on the single measurement is given by the combination in quadrature of the individual and the common standard uncertainties, and all the covariances are equal to σ_c^2. To verify, in a simple case, that the result is reasonable, let us consider only two independent quantities X_1 and X_2, and a calibration constant $X_3 = c$, having an expected value equal to zero. From these we can calculate the correlated quantities Y_1 and Y_2 and finally their sum ($S \equiv Z_1$) and difference ($D \equiv Z_2$). The results are

$$\mathbf{V}_Y = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix}, \tag{6.31}$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 + 4\,\sigma_c^2 & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}. \tag{6.32}$$

It follows that

$$\begin{aligned}
\sigma^2(S) &= \sigma_1^2 + \sigma_2^2 + (2\,\sigma_c)^2, & (6.33)\\
\sigma^2(D) &= \sigma_1^2 + \sigma_2^2\,, & (6.34)
\end{aligned}$$

as intuitively expected.

Normalization uncertainty

Let us consider now the case where the calibration constant is the scale factor f, known with a standard uncertainty σ_f. Also in this case, for simplicity and without losing generality, let us suppose that the most probable value of f is 1. Then $X_{n+1} = f$, i.e. $x_{n+1} = 1$, and $V_{X_{n+1,n+1}} = \sigma_f^2$. Then

$$\sigma^2(Y_i) = \sigma_i^2 + \sigma_f^2\, x_i^2\,, \tag{6.35}$$

$$\mathrm{Cov}(Y_i, Y_j) = \sigma_f^2\, x_i\, x_j \qquad (i \neq j)\,, \tag{6.36}$$

$$\rho_{ij} = \frac{x_i\, x_j}{\sqrt{x_i^2 + \frac{\sigma_i^2}{\sigma_f^2}}\sqrt{x_j^2 + \frac{\sigma_j^2}{\sigma_f^2}}}\,, \tag{6.37}$$

$$|\rho_{ij}| = \frac{1}{\sqrt{1 + \left(\frac{\sigma_i}{\sigma_f\, x_i}\right)^2}\sqrt{1 + \left(\frac{\sigma_j}{\sigma_f\, x_j}\right)^2}}\,. \tag{6.38}$$

To verify the results let us consider two independent measurements X_1 and X_2; let us calculate the correlated quantities Y_1 and Y_2, and finally their product ($P \equiv Z_1$) and their ratio ($R \equiv Z_2$):

$$\mathbf{V}_Y = \begin{pmatrix} \sigma_1^2 + \sigma_f^2\, x_1^2 & \sigma_f^2\, x_1\, x_2 \\ \sigma_f^2\, x_1\, x_2 & \sigma_2^2 + \sigma_f^2\, x_2 \end{pmatrix}, \tag{6.39}$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2\, x_2^2 + \sigma_2^2\, x_1^2 + 4\,\sigma_f^2\, x_1^2\, x_2^2 & \sigma_1^2 - \sigma_2^2\, \frac{x_1^2}{x_2^2} \\ \sigma_1^2 - \sigma_2^2\, \frac{x_1^2}{x_2^2} & \frac{\sigma_1^2}{x_2^2} + \sigma_2^2\, \frac{x_1^2}{x_2^4} \end{pmatrix}. \tag{6.40}$$

It follows that

$$\sigma^2(P) = \sigma_1^2\, x_2^2 + \sigma_2^2\, x_1^2 + (2\,\sigma_f\, x_1\, x_2)^2\,, \tag{6.41}$$

$$\sigma^2(R) = \frac{\sigma_1^2}{x_2^2} + \sigma_2^2\, \frac{x_1^2}{x_2^4}\,. \tag{6.42}$$

Just as an unknown common offset error cancels in differences and is enhanced in sums, an unknown normalization error has a similar effect on the ratio and the product. It is also interesting to calculate the standard uncertainty of a difference in the case of a normalization error:

$$\sigma^2(D) = \sigma_1^2 + \sigma_2^2 + \sigma_f^2\,(x_1 - x_2)^2\,. \tag{6.43}$$

The contribution from an unknown normalization error vanishes if the two values are equal.

General case

Let us assume there are n independently measured values x_i and m calibration constants c_j with their covariance matrix $\mathbf{V}_c$. The latter can also be theoretical parameters influencing the data, and moreover they may be correlated, as usually happens if, for example, they are parameters of a calibration fit. We can then include the c_j in the vector that contains the measurements

and $\mathbf{V}_c$ in the covariance matrix $\mathbf{V}_X$:

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ c_1 \\ \vdots \\ c_m \end{pmatrix}, \qquad \mathbf{V}_X = \left(\begin{array}{cccc|c} \sigma_1^2 & 0 & \cdots & 0 & \\ 0 & \sigma_2^2 & \cdots & 0 & \\ \cdots & \cdots & \cdots & \cdots & \mathbf{0} \\ 0 & 0 & \cdots & \sigma_n^2 & \\ \hline & & \mathbf{0} & & \mathbf{V}_c \end{array}\right). \tag{6.44}$$

The corrected quantities are obtained from the most general function

$$Y_i = Y_i(X_i, \underline{c}) \qquad\qquad (i = 1, 2, \ldots, n), \tag{6.45}$$

and the covariance matrix $\mathbf{V}_Y$ from the covariance propagation $\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^T$.

As a frequently encountered example, we can think of several normalization constants, each affecting a subsample of the data – as is the case where each of several detectors measures a set of physical quantities. Let us consider just three quantities (X_i) and three uncorrelated normalization standard uncertainties (σ_{f_j}), the first common to X_1 and X_2, the second to X_2 and X_3 and the third to all three. We get the following covariance matrix:

$$\begin{pmatrix} \sigma_1^2 + \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1^2 & \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1 x_2 & \sigma_{f_3}^2 x_1 x_3 \\ \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1 x_2 & \sigma_2^2 + \left(\sigma_{f_1}^2 + \sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2^2 & \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2 x_3 \\ \sigma_{f_3}^2 x_1 x_3 & \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2 x_3 & \sigma_3^2 + \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_3^2 \end{pmatrix}. \tag{6.46}$$

6.3.2 Use and misuse of the covariance matrix to fit correlated data

Best estimate of the true value from two correlated values.

Once the covariance matrix is built one uses it in a χ^2 fit to get the parameters of a function. The quantity to be minimized is χ^2, defined as

$$\chi^2 = \underline{\Delta}^T \mathbf{V}^{-1} \underline{\Delta}, \tag{6.47}$$

where $\underline{\Delta}$ is the vector of the differences between the theoretical and the experimental values. Let us consider the simple case in which two results of the same physical quantity are available, and the individual and the common standard uncertainty are known. The best estimate of the true value of the physical quantity is then obtained by fitting the constant $Y = k$ through the data points. In this simple case the χ^2 minimization can be performed easily. We will consider the two cases of offset and normalization uncertainty. As before, we assume that the detector is well calibrated, i.e. the most probable value of the calibration constant is, respectively for the two cases, 0 and 1, and hence $y_i = x_i$.

Offset uncertainty

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_c the common standard uncertainty:

$$\begin{aligned} \chi^2 \;=\; & \frac{1}{D}\left[(x_1 - k)^2\,(\sigma_2^2 + \sigma_c^2) + (x_2 - k)^2\,(\sigma_1^2 + \sigma_c^2)\right. \\ & \left. -2\,(x_1 - k)\,(x_2 - k)\,\sigma_c^2\right], \end{aligned} \tag{6.48}$$

where $D = \sigma_1^2\,\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\,\sigma_c^2$ is the determinant of the covariance matrix.

Minimizing χ^2 and using the second derivative calculated at the minimum we obtain the best value of k and its standard deviation:

$$\widehat{k} \;=\; \frac{x_1\,\sigma_2^2 + x_2\,\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (=\overline{x}), \tag{6.49}$$

$$\sigma^2(\widehat{k}) \;=\; \frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \sigma_c^2\,. \tag{6.50}$$

The most probable value of the physical quantity is exactly that which one obtains from the average $\overline{x}$ weighted with the inverse of the individual variances. Its overall uncertainty is the quadratic sum of the standard deviation of the weighted average and the common one. The result coincides with the simple expectation.

Normalization uncertainty

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_f the common standard uncertainty on the scale:

$$\begin{aligned}\chi^2 \;=\; \frac{1}{D}\,\big[&(x_1-k)^2\,(\sigma_2^2 + x_2^2\,\sigma_f^2) + (x_2-k)^2\,(\sigma_1^2 + x_1^2\,\sigma_f^2)\\ &-2\cdot(x_1-k)\cdot(x_2-k)\cdot x_1\cdot x_2\cdot\sigma_f^2\big]\,,\end{aligned} \tag{6.51}$$

where $D = \sigma_1^2\,\sigma_2^2 + (x_1^2\,\sigma_2^2 + x_2^2\,\sigma_1^2)\,\sigma_f^2$. We obtain in this case the following result:

$$\widehat{k} \;=\; \frac{x_1\,\sigma_2^2 + x_2\,\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\,\sigma_f^2}, \tag{6.52}$$

$$\sigma^2(\widehat{k}) \;=\; \frac{\sigma_1^2\,\sigma_2^2 + (x_1^2\,\sigma_2^2 + x_2^2\,\sigma_1^2)\,\sigma_f^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\,\sigma_f^2}\,. \tag{6.53}$$

With respect to the previous case, $\widehat{k}$ has a new term $(x_1 - x_2)^2\,\sigma_f^2$ in the denominator. As long as this is negligible with respect to the individual variances we still get the weighted average $\overline{x}$, otherwise a smaller value is obtained. Calling r the ratio between $\widehat{k}$ and $\overline{x}$, we obtain

$$r = \frac{\widehat{k}}{\overline{x}} = \frac{1}{1 + \frac{(x_1-x_2)^2}{\sigma_1^2+\sigma_2^2}\,\sigma_f^2}\,. \tag{6.54}$$

Written in this way, one can see that the deviation from the simple average value depends on the compatibility of the two values and on the normalization uncertainty. This can be understood in the following way: as soon as the two values are in some disagreement, the fit starts to vary the normalization factor (in a hidden way) and to squeeze the scale by an amount allowed by σ_f, in order to minimize the χ^2. The reason the fit prefers normalization factors smaller than 1 under these conditions lies in the standard formalism of the covariance propagation, where only first derivatives are considered. This implies that the individual standard deviations are not rescaled by lowering the normalization factor, but the points get closer.

Example 1. Consider the results of two measurements, 8.0 and 8.5, having 2% individual and 10% common normalization uncertainty. Assuming that the two measurements refer to the same physical quantity, the best estimate of its true value can be obtained by fitting

the points to a constant. Minimizing χ^2 with $\mathbf{V}$ estimated empirically by the data, as explained in the previous section, one obtains a value of 7.87 ± 0.81, which is surprising to say the least, since the most probable result is outside the interval determined by the two measured values.

Example 2. A real life case of this strange effect which occurred during the global analysis of the R ratio in e^+e^- performed by The CELLO Collaboration [48], is shown in Fig. 6.1. The data points represent the averages in energy bins of the results of the PETRA and PEP experiments. They are all correlated and the bars show the total uncertainty (see Refs. [48] and [49] for details). In particular, at the intermediate stage of the analysis shown in the figure, an overall 1% systematic error due theoretical uncertainties was included in the covariance matrix. The R values above 36 GeV show the first hint of the rise of the e^+e^- cross-section due to the Z° pole. At that time it was very interesting to prove that the observation was not just a statistical fluctuation. In order to test this, the R measurements were fitted with a theoretical function having no Z° contributions, using only data below a certain energy. It was expected that a fast increase of χ^2 per number of degrees of freedom ν would be observed above 36 GeV, indicating that a theoretical prediction without Z° would be inadequate for describing the high-energy data. The surprising result was a repulsion (see Fig. 6.1) between the experimental data and the fit: Including the high-energy points with larger R a lower curve was obtained, while χ^2/ν remained almost constant.

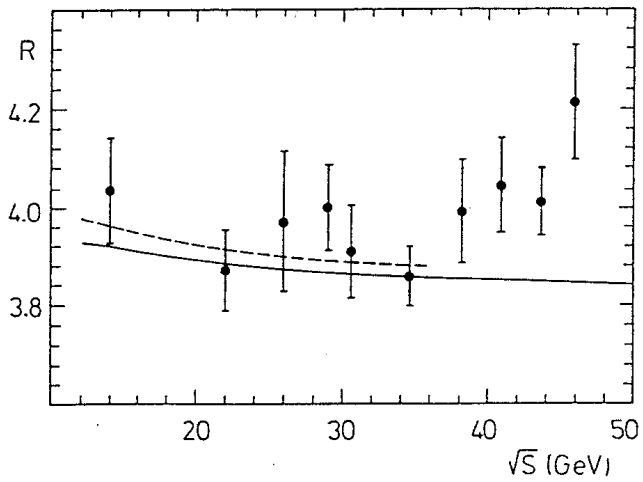

Figure 6.1: R measurements from PETRA and PEP experiments with the best fits of QED+QCD to all the data (full line) and only below 36 GeV (dashed line). All data points are correlated (see text).

To see the source of this effect more explicitly let us consider an alternative way often used to take the normalization uncertainty into account. A scale factor f, by which all data points are multiplied, is introduced to the expression of the χ^2:

$$\chi^2_A = \frac{(f\,x_1 - k)^2}{(f\,\sigma_1)^2} + \frac{(f\,x_2 - k)^2}{(f\,\sigma_2)^2} + \frac{(f-1)^2}{\sigma_f^2}\,. \tag{6.55}$$

Let us also consider the same expression when the individual standard deviations are not rescaled:

$$\chi^2_B = \frac{(f\,x_1 - k)^2}{\sigma_1^2} + \frac{(f\,x_2 - k)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}\,. \tag{6.56}$$

The use of χ^2_A always gives the result $\widehat{k} = \overline{x}$, because the term $(f-1)^2/\sigma_f^2$ is harmless[9] as far as the value of the minimum χ^2 and the determination on $\widehat{k}$ are concerned. Its only influence is on $\sigma(\widehat{k})$, which turns out to be equal to quadratic combination of the weighted average standard deviation with $\sigma_f\,\overline{x}$, the normalization uncertainty on the average. This result corresponds to the usual one when the normalization factor in the definition of χ^2 is not included, and the overall uncertainty is added at the end.

Instead, the use of χ^2_B is equivalent to the covariance matrix: The same values of the minimum χ^2, of $\widehat{k}$ and of $\sigma(\widehat{k})$ are obtained, and $\widehat{f}$ at the minimum turns out to be exactly the r ratio defined above. This demonstrates that the effect happens when the data values are rescaled independently of their standard uncertainties. The effect can become huge if the data show mutual disagreement. The equality of the results obtained with χ^2_B with those obtained with the covariance matrix allows us to study, in a simpler way, the behaviour of r $(= \widehat{f})$ when an arbitrary number of data points are analysed. The fitted value of the normalization factor is

$$\widehat{f} = \frac{1}{1 + \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_i^2}\,\sigma_f^2}\,. \tag{6.57}$$

If the values of x_i are consistent with a common true value it can be shown that the expected value of $\widehat{f}$ is

$$\langle \widehat{f} \rangle \;= \frac{1}{1 + (n-1)\,\sigma_f^2}\,. \tag{6.58}$$

Hence, there is a bias on the result when for a non-vanishing σ_f a large number of data points are fitted. In particular, the fit on average produces a bias larger than the normalization uncertainty itself if $\sigma_f > 1/(n-1)$. One can also see that $\sigma^2(\widehat{k})$ and the minimum of χ^2 obtained with the covariance matrix or with χ^2_B are smaller by the same factor r than those obtained with χ^2_A.

Peelle's Pertinent Puzzle

To summarize, when there is an overall uncertainty due to an unknown systematic error and the covariance matrix is used to define χ^2, the behaviour of the fit depends on whether the uncertainty is on the offset or on the scale. In the first case the best estimates of the function parameters are exactly those obtained without overall uncertainty, and only the parameters' standard deviations are affected. In the case of unknown normalization errors, biased results can be obtained. The size of the bias depends on the fitted function, on the magnitude of the overall uncertainty and on the number of data points.

It has also been shown that this bias comes from the linearization performed in the usual covariance propagation. This means that, even though the use of the covariance matrix can be

[9] This can be seen by rewriting (6.55) as

$$\frac{(x_1 - k/f)^2}{\sigma_1^2} + \frac{(x_2 - k/f)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}\,.$$

For any f, the first two terms determine the value of k, and the third one binds f to 1.

very useful in analysing the data in a compact way using available computer algorithms, care is required if there is one large normalization uncertainty which affects all the data.

The effect discussed above has also been observed independently by R.W. Peelle and reported the year after the analysis of the CELLO data [48]. The problem has been extensively discussed among the community of nuclear physicists, where it is currently known as 'Peelle's Pertinent Puzzle' [50].

Recent cases in HEP in which this effect has been found to have biased the result are discussed in Refs. [51, 52].

Note added: the solution outlined here is taken from Ref. [47], and it has to be considered an *ad hoc* solution. The general (of course Bayesian) solution to the χ^2 paradox has been worked out recently [53], and it will be published in a forthcoming paper.

Chapter 7

Bayesian unfolding

"Now we see but a poor reflection as in a mirror ... "
"Now I know in part ... "
(1 Cor.)

7.1 Problem and typical solutions

In any experiment the distribution of the measured observables differs from that of the corresponding true physical quantities due to physics and detector effects. For example, one may be interested in measuring the variables x and Q^2 in deep-inelastic scattering events. In such a case one is able to build statistical estimators which in principle have a physical meaning similar to the true quantities, but which have a non-vanishing variance and are also distorted due to QED and QCD radiative corrections, parton fragmentation, particle decay and limited detector performances. The aim of the experimentalist is to unfold the observed distribution from all these distortions so as to extract the true distribution (see also Refs. [54] and [55]). This requires a satisfactory knowledge of the overall effect of the distortions on the true physical quantity.

When dealing with only one physical variable the usual method for handling this problem is the so-called 'bin-to-bin' correction: one evaluates a generalized efficiency (it may even be larger than unity) by calculating the ratio between the number of events falling in a certain bin of the reconstructed variable and the number of events in the same bin of the true variable with a Monte Carlo simulation. This efficiency is then used to estimate the number of true events from the number of events observed in that bin. Clearly this method requires the same subdivision in bins of the true and the experimental variable and hence it cannot take into account large migrations of events from one bin to the others. Moreover it neglects the unavoidable correlations between adjacent bins. This approximation is valid only if the amount of migration is negligible and if the standard deviation of the smearing is smaller than the bin size.

An attempt to solve the problem of migrations is sometimes made by building a matrix which connects the number of events generated in one bin to the number of events observed in the other bins. This matrix is then inverted and applied to the measured distribution. This immediately produces inversion problems if the matrix is singular. On the other hand, there is no reason from a probabilistic point of view why the inverse matrix should exist. This can easily be seen by taking the example of two bins of the true quantity both of which have the same probability of being observed in each of the bins of the measured quantity. It follows that treating probability distributions as vectors in space is not correct, even in principle. Moreover the method is not able to handle large statistical fluctuations even if the matrix can be inverted

(if we have, for example, a very large number of events with which to estimate its elements and we choose the binning in such a way as to make the matrix not singular). The easiest way to see this is to think of the unavoidable negative terms of the inverse of the matrix which in some extreme cases may yield negative numbers of unfolded events. Quite apart from these theoretical reservations, the actual experience of those who have used this method is rather discouraging, the results being highly unstable.

7.2 Bayes' theorem stated in terms of causes and effects

Let us state Bayes' theorem in terms of several independent causes ($C_i,\ i = 1, 2, \ldots, n_C$) which can produce one effect (E). For example, if we consider deep-inelastic scattering events, the effect E can be the observation of an event in a cell of the measured quantities $\{\Delta Q^2_{meas}, \Delta x_{meas}\}$. The causes C_i are then all the possible cells of the true values $\{\Delta Q^2_{true}, \Delta x_{true}\}_i$. Let us assume we know the initial probability of the causes $P(C_i)$ and the conditional probability that the i-th cause will produce the effect $P(E \,|\, C_i)$. The Bayes formula is then

$$P(C_i \,|\, E) = \frac{P(E \,|\, C_i)\, P(C_i)}{\sum_{l=1}^{n_C} P(E \,|\, C_l)\, P(C_l)}\,. \tag{7.1}$$

$P(C_i \,|\, E)$ depends on the initial probability of the causes. If one has no better prejudice concerning $P(C_i)$ the process of inference can be started from a uniform distribution.

The final distribution depends also on $P(E \,|\, C_i)$. These probabilities must be calculated or estimated with Monte Carlo methods. One has to keep in mind that, in contrast to $P(C_i)$, these probabilities are not updated by the observations. So if there are ambiguities concerning the choice of $P(E \,|\, C_i)$ one has to try them all in order to evaluate their systematic effects on the results.

7.3 Unfolding an experimental distribution

If one observes $n(E)$ events with effect E, the expected number of events assignable to each of the causes is

$$\widehat{n}(C_i) = n(E)\, P(C_i \,|\, E)\,. \tag{7.2}$$

As the outcome of a measurement one has several possible effects E_j ($j = 1, 2, \ldots, n_E$) for a given cause C_i. For each of them the Bayes formula (7.1) holds, and $P(C_i \,|\, E_j)$ can be evaluated. Let us write (7.1) again in the case of n_E possible effects,[1] indicating the initial probability of the causes with $P_\circ(C_i)$:

$$P(C_i \,|\, E_j) = \frac{P(E_j \,|\, C_i)\, P_\circ(C_i)}{\sum_{l=1}^{n_C} P(E_j \,|\, C_l)\, P_\circ(C_l)}\,. \tag{7.3}$$

One should note the following.

- $\sum_{i=1}^{n_C} P_\circ(C_i) = 1$, as usual. Note that if the probability of a cause is initially set to zero it can never change, i.e. if a cause does not exist it cannot be invented.

[1] The broadening of the distribution due to the smearing suggests a choice of n_E larger than n_C. It is worth mentioning that there is no need to reject events where a measured quantity has a value outside the range allowed for the physical quantity. For example, in the case of deep-inelastic scattering events, cells with $x_{meas} > 1$ or $Q^2_{meas} < 0$ give information about the true distribution too.

- $\sum_{i=1}^{n_C} P(C_i \,|\, E_j) = 1$. This normalization condition, mathematically trivial since it comes directly from (7.3), indicates that each effect must come from one or more of the causes under examination. This means that if the observables also contain a non-negligible amount of background, this needs to be included among the causes.

- $0 \leq \epsilon_i \equiv \sum_{j=1}^{n_E} P(E_j \,|\, C_i) \leq 1$. There is no need for each cause to produce at least one of the effects. ϵ_i gives the efficiency of finding the cause C_i in any of the possible effects.

After N_{obs} experimental observations one obtains a distribution of frequencies $\underline{n}(E) \equiv \{n(E_1), n(E_2), \ldots, n(E_{n_E})\}$. The expected number of events to be assigned to each of the causes (taking into account only the observed events) can be calculated by applying (7.2) to each effect:

$$\widehat{n}(C_i)|_{obs} \;=\; \sum_{j=1}^{n_E} n(E_j)\, P(C_i \,|\, E_j)\,. \tag{7.4}$$

When inefficiency[2] is also brought into the picture, the best estimate of the true number of events becomes

$$\widehat{n}(C_i) \;=\; \frac{1}{\epsilon_i} \sum_{j=1}^{n_E} n(E_j)\, P(C_i \,|\, E_j) \qquad \epsilon_i \neq 0\,. \tag{7.5}$$

From these unfolded events we can estimate the true total number of events, the final probabilities of the causes and the overall efficiency:

$$\begin{aligned} \widehat{N}_{true} &= \sum_{i=1}^{n_C} \widehat{n}(C_i)\,, \\ \widehat{P}(C_i) \equiv P(C_i \,|\, \underline{n}(E)) &= \frac{\widehat{n}(C_i)}{\widehat{N}_{true}}\,, \\ \widehat{\epsilon} &= \frac{N_{obs}}{\widehat{N}_{true}}\,. \end{aligned}$$

If the initial distribution $\underline{P_\circ}(C)$ is not consistent with the data, it will not agree with the final distribution $\underline{\widehat{P}}(C)$. The closer the initial distribution is to the true distribution, the better the agreement is. For simulated data one can easily verify that the distribution $\underline{\widehat{P}}(C)$ lies between $\underline{P_\circ}(C)$ and the true one. This suggests proceeding iteratively. Figure 7.1 shows an example of a bidimensional distribution unfolding.

More details about iteration strategy, evaluation of uncertainty, etc. can be found in Ref. [56]. I would just like to comment on an obvious criticism that may be made: 'the iterative procedure is against the Bayesian spirit, since the same data are used many times for the same inference'. In principle the objection is valid, but in practice this technique is a trick to give to the experimental data a weight (an importance) larger than that of the priors. A more rigorous procedure which took into account uncertainties and correlations of the initial distribution would have been much more complicated. An attempt of this kind can be found in Ref. [57]. Examples of unfolding procedures performed with non-Bayesian methods are described in Refs. [54] and [55].

Note added: A recent book by Cowan [58] contains an interesting chapter on unfolding. More sophisticated methods for, generally speaking, image reconstruction can be found in Ref. [59] and references therein.

[2] If $\epsilon_i = 0$ then $\widehat{n}(C_i)$ will be set to zero, since the experiment is not sensitive to the cause C_i.

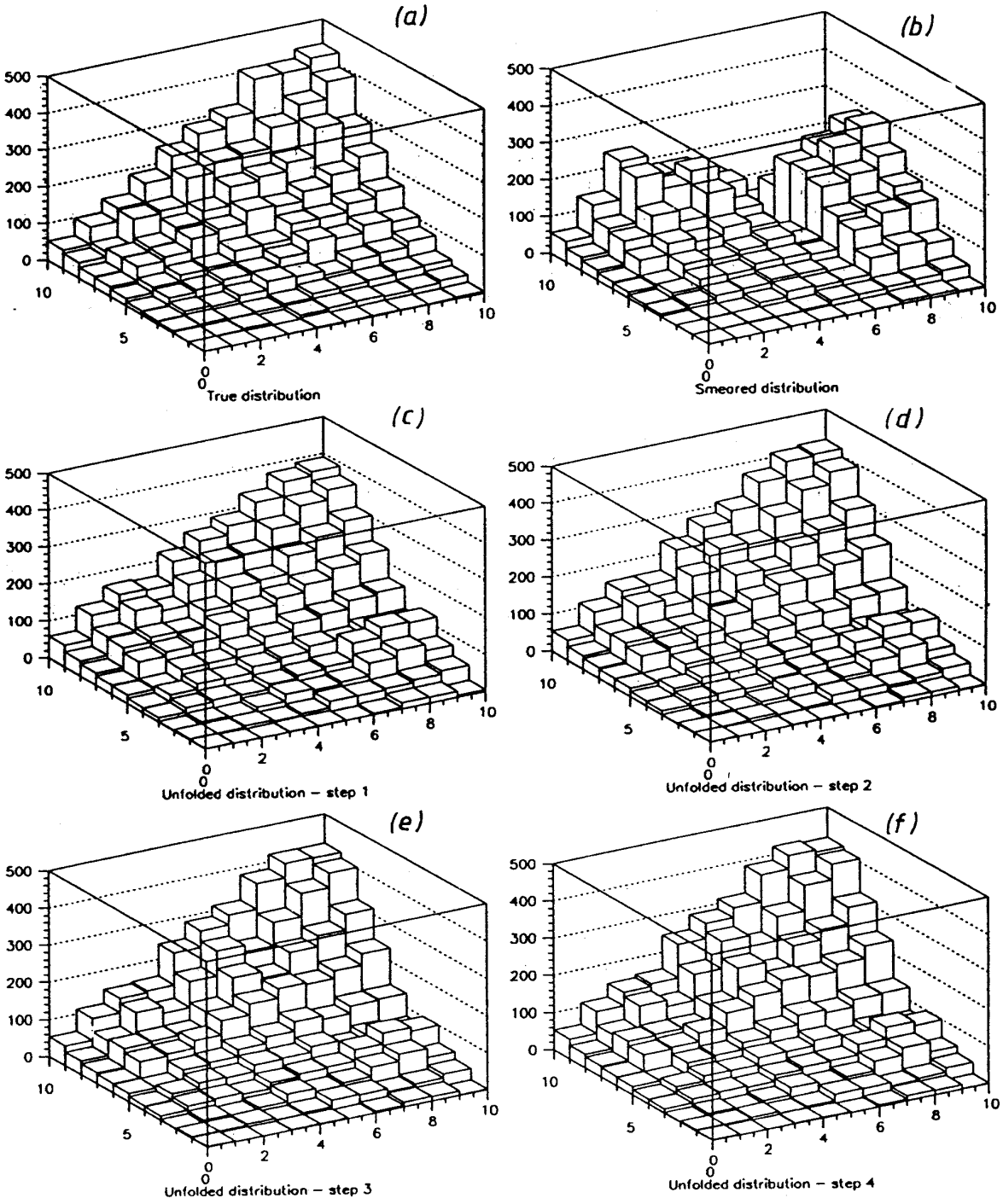

Figure 7.1: Example of two-dimensional unfolding: true distribution (a), smeared distribution (b) and results after the first four steps [(c) to (f)].

Part III

Other comments, examples and applications

Chapter 8

Appendix on probability and inference

8.1 Unifying role of subjective approach

I would like to give some examples to clarify what I mean by 'linguistic schizophrenia' (see Section 3.3.2). Let us consider the following:

1. probability of a '6' when tossing a die;
2. probability that the 100001st event will be accepted in the cuts of the analysis of simulated events, if I know that 91 245 out of 100 000 events[1] have already been accepted;
3. probability that a real event will be accepted in the analysis, given the knowledge of point 2, and assuming that exactly the same analysis program is used, and that the Monte Carlo describes best the physics and the detector;
4. probability that an observed track is π^+, if I have learned from the Monte Carlo that ... ;
5. probability that the Higgs mass is greater than 400 GeV;
6. probability that the 1000th decimal digit of π is 5;
7. probability of rain tomorrow;
8. probability that the US dollar will be exchanged at ≥ 2 DM before the end of 1999 (statement made in spring 1998).

Let us analyse in detail the statements.

- The evaluation of point 1 is based on considerations of physical symmetry, using the combinatorial evaluation rule. The first remark is that a convinced frequentist should abstain from assessing such a probability until he has collected statistical data on that die. Otherwise he is implicitly assuming that the frequency-based definition is not a definition, but one of the possible evaluation rules (and then the concept can only be that related to the degree of belief ...).

[1] Please note that 'event' is also used here according to HEP jargon (this is quite a case of homonymy to which one has to pay attention, but it has nothing to do with the linguistic schizophrenia I am talking about).

For those who, instead, believe that probability is only related to symmetry the answer appears to be absolutely objective: 1/6. But it is clear that one is in fact giving a very precise and objective answer to something that is not real. Instead, we should only talk about reality. This example should help to clarify the de Finetti sentence quoted in Section 2.2 (*"The classical view ... "*, in particular, *"The original sentence becomes meaningful if reversed ... "*).

- Point 2 leads to a consistent answer within the frequentistic approach, which is numerically equal to the subjective one [see, for example, (5.33) and (5.36)], whilst it has no solution in a combinatorial definition.
- Points 3 and 4 are different from point 2. The frequentistic definition is not applicable. The translation from simulated events to real events is based on beliefs, which may be as firmly based as you like, but they remain beliefs. So, although this operation is routinely carried out by every experimentalist, it is meaningful only if the probability is meant as a degree of belief and not a limit of relative frequency.
- Points 3–7 are only meaningful if probability is interpreted as a degree of belief.[2]

The unifying role of subjective probability should be clear from these examples. All those who find statements 1–7 meaningful, are implicitly using subjective probability. If not, there is nothing wrong with them, on condition that they make probabilistic statements only in those cases where their definition of probability is applicable (essentially never in real life and in research). If, however, they still insist on speaking about probability outside the condition of validity of their definition, refusing the point of view of subjective probability, they fall into the self-declared linguistic schizophrenia of which I am talking, and they generate confusion.[3]

Another very important point is the crucial role of coherence (see Section 3.3.2), which allows the exchange of the value of the probability between rational individuals: if someone tells me that he judges the probability of a given event to be 68%, then I imagine that he is as confident about it as he would be about extracting a white ball from a box which contains 100 balls, 68 of which are white. This event could be related, for example, to the result of a measurement:

$$\mu = \mu_\circ \pm \sigma(\mu)\,,$$

assuming a Gaussian model. If an experimentalist feels ready to place a 2:1 bet[4] in favour of the statement, but not a 1:2 bet against it, it means that his assessment of probability is not coherent. In other words, he is cheating, for he knows that his result will be interpreted differently from what he really believes (he has consciously overestimated the 'error bar', because he is afraid of being contradicted). If you want to know whether a result is coherent, take an interval given by 70% of the quoted uncertainty and ask the experimentalist if he is ready to place a 1:1 bet in either direction.

8.2 Frequentists and combinatorial evaluation of probability

In the previous section it was was said that frequentists should abstain from assessing probabilities if a long-run experiment has not been carried out. But frequentists do, using a sophisticated

[2] In fact, one could use the combinatorial evaluation in point 6 as well, because of the discussed cultural reasons, but not everybody is willing to speak about the probability of something which has a very precise value, although unknown.

[3] See for example Refs. [46] and [60], where it is admitted that the Bayesian approach is good for decision problems, although they stick to the frequentistic approach.

[4] This corresponds to a probability of $2/3 \approx 68\%$.

reasoning, of which perhaps not everyone is aware. I think that the best way to illustrate this reasoning is with an example of an authoritative exponent, Polya [61], who adheres to von Mises' views [62].

> *"A bag contains p balls of various colors among which there are exactly f white balls. We use this simple apparatus to produce a random mass phenomenon. We draw a ball, we look at its color and we write W if the ball is white, but we write D if it is of a different color. We put back the ball just drawn into the bag, we shuffle the balls in the bag, then we draw again one and note the color of this second ball, W or D. In proceeding so, we obtain a random sequence (...):*
>
> $$W\ D\ D\ D\ W\ D\ D\ W\ W\ D\ D\ D\ W\ W\ D\ .$$
>
> *What is the long range relative frequency of the white balls?*
>
> *Let us assume that the balls are homogeneous and exactly spherical, made of the same material and having the same radius. Their surfaces are equally smooth, and their different coloration influences only negligibly their mechanical behavior, if it has any influence at all. The person who draws the balls is blindfolded or prevented in some other manner from seeing the balls. The position of the balls in the bag varies from one drawing to the other, is unpredictable, beyond our control. Yet the permanent circumstances are well under control: the balls are all the same shape, size, and weight; they are indistinguishable by the person who draws them.*
>
> *Under such circumstances we see no reason why one ball should be preferred to another and we naturally expect that, in the long run, each ball will be drawn approximately equally often. Let us say that we have the patience to make 10 000 drawings. Then we should expect that each of the p balls will appear about*
>
> $$\frac{10\,000}{p}\ times.$$
>
> *There are f white balls. Therefore, in 10 000 drawings, we expect to get white*
>
> $$f\,\frac{10\,000}{p} = 10\,000\,\frac{f}{p}\ times;$$
>
> *this is the expected frequency of the white balls. To obtain the relative frequency, we have to divide by the number of observations, or drawings, that is, 10 000. And so we are led to the statement: the long range relative frequency, or probability, of the white balls is f/p.*
>
> *The letters f and p are chosen to conform to the traditional mode of expression. As we have to draw one of the p balls, we have to choose one of p possible cases. We have good reasons (equal condition of the p balls) not to prefer any of these p possible cases to any other. If we wish that a white ball should be drawn (for example, if we are betting on white), the f white balls appear to us as favourable cases. Hence we can describe the probability f/p as the ratio of the number of favourable cases to the number of possible cases."*

The approach sketched in the above example is based on the refusal of calling probability (the intuitive concept of it) by its name. The term 'probability' is used instead for 'long-range relative frequency'. Nevertheless, the value of probability is not evaluated from the information about past frequency, but from the hypothetical long-range relative frequency, based on: a) plausible (and subjective!) reasoning on equiprobability (although not stated with this term) of the possible outcomes; b) the expectation ($\equiv$belief) that the relative frequency will be equal to the fraction of white balls in the bag.[5] The overall result is to confuse the matter, without any

[5] Sometimes this expectation is justified advocating the law of large numbers, expressed by the Bernoulli theorem. This is unacceptable, as pointed out by de Finetti: *"For those who seek to connect the notion of*

philosophical or practical advantages (compare the twisted reasoning of the above example with Hume's lucid exposure of the concept of probability and its evaluation by symmetry arguments, reported in Section 2.2).

8.3 Interpretation of conditional probability

As repeated throughout these notes, and illustrated with many examples, probability is always conditioned probability. Absolute probability makes no sense. Nevertheless, there is still something in the primer which can be misleading and that needs to be clarified, namely the so-called 'formula of conditional probability' (Section 3.4.2):

$$P(E\,|\,H) = \frac{P(E \cap H)}{P(H)} \qquad (P(H) \neq 0)\,. \tag{8.1}$$

What does it mean? Textbooks present it as a definition (a kind of 4th axiom), although very often, a few lines later in the same book, the formula $P(E \cap H) = P(E\,|\,H) \cdot P(H)$ is presented as a theorem (!).

In the subjective approach, one is allowed to talk about $P(E\,|\,H)$ independently of $P(E \cap H)$ and $P(H)$. In fact, $P(E\,|\,H)$ is just the assessment of the probability of E, under the condition that H is true. Then it cannot depend on the probability of H. It is easy to show with an example that this point of view is rather natural, whilst that of considering (8.1) as a definition is artificial. Let us take

- H = Higgs mass of 250 GeV;
- E = the decay products which are detected in a LHC detector;
- the evaluation of $P(E\,|\,H)$ is a standard PhD student task. He chooses $m_H = 250\,\mathrm{GeV}$ in the Monte Carlo and counts how many events pass the cuts (for the interpretation of this operation, see the previous section). No one would think that $P(E\,|\,H)$ must be evaluated only from $P(E \cap H)$ and $P(H)$, as the definition (8.1) would imply. Moreover, the procedure is legitimate even if we knew with certainty that the Higgs mass was below 200 GeV and, therefore, $P(H) = 0$.

In the subjective approach, (8.1) is a true theorem required by coherence. It means that although one can speak of each of the three probabilities independently of the others, once two of them have been elicited, the third is constrained. It is interesting to demonstrate the theorem to show that it has nothing to do with the kind of heuristic derivation of Section 3.4.2:

- Let us imagine a coherent bet on the conditional event $E\,|\,H$ to win a unitary amount of money ($B = 1$, as the scale factor is inessential). Remembering the meaning of conditional probability in terms of bets (see Section 3.4.2), this means that
  - we pay (with certainty) $A = P(E\,|\,H)$;
  - we win 1 if E and H are both verified (with probability $P(E \cap H)$);

probability with that of frequency, results which relate probability and frequency in some way (and especially those results like the 'law of large numbers') play a pivotal rôle, providing support for the approach and for the identification of the concepts. Logically speaking, however, one cannot escape from the dilemma posed by the fact that the same thing cannot both be assumed first as a definition and then proved as a theorem; nor can one avoid the contradiction that arises from a definition which would assume as certain something that the theorem only states to be very probable."[11]

– we get our money back (i.e. A) if H does not happen (with probability $P(\overline{H})$).

- The expected value of the 'gain' G is given by the probability of each event multiplied by the gain associated with each event:

$$\mathrm{E}(G) \;=\; 1\cdot(-P(E\,|\,H)) + P(E\cap H)\cdot 1 + P(\overline{H})\cdot P(E\,|\,H)\,,$$

where the first factors of the products on the right-hand side of the formula stand for probability, the second for the amount of money. It follows that

$$\begin{aligned}\mathrm{E}(G) &= -P(E\,|\,H) + P(E\cap H) + (1-P(H))\cdot P(E\,|\,H)\\ &= P(E\cap H) - P(E\,|\,H)\cdot P(H)\,. \end{aligned} \tag{8.2}$$

- Coherence requires the rational better to be indifferent to the direction of the bet, i.e. $\mathrm{E}(G)=0$. Applying this condition to (8.2) we obtain (8.1).

8.4 Are the beliefs in contradiction to the perceived objectivity of physics?

This is one of the most important points to be clarified since it is felt by many to be the biggest obstacle, preventing them from understanding the Bayesian approach: is there a place for beliefs in science? The usual criticism is that science must be objective and, hence, that there should be no room for subjectivity. A colleague once told me: *"I do not believe something. I assess it. This is not a matter for religion!"*

As I understand it, there are two possible ways to surmount the obstacle. The first is to try to give a more noble status of objectivity to the Bayesian approach, for example by formulating objective priors. In my opinion the main result of this attempt is to spoil the original nature of the theory, by adding dogmatic ingredients [22]. The second way consists, more simply, in recognizing that beliefs are a natural part of doing science. Admitting that they exist does not spoil the perceived objectivity of well-established science. In other words, one needs only to look closely at how frontier science makes progress, instead of seeking refuge in an idealized concept of objectivity.[6]

Clearly this discussion would require another book, and not just some side remarks, but I am confident that the reader for whom this report is intended, and who is supposed to have working experience in frontier research, is already prepared for what I am going to say. I find it hard to discuss these matters with people who presume to teach us about the way physics, and science in general, proceeds, without having the slightest direct experience of what they are talking about.

First of all, I would like to invite you to pay attention to the expressions we use in private and public discussions, and in written matter too.[7] Here are some examples:

- "I believe that ... ";
- "We have to get experience with ... ";

[6] My preferred motto on this matter is *"no one should be allowed to speak about objectivity unless he has had 10–20 years working experience in frontier science, economics, or any other applied field"*.

[7] For example, the statistician D. Berry [63] has amused himself by counting how many times Hawking uses 'belief', 'to believe', or synonyms, in his *'A brief history of time'*. The book could have been entitled *'A brief history of beliefs'*, pointed out Berry in his talk ...

- "I don't trust that guy (or that collaboration, or that procedure)";
- "Oh yes, if this has been told you by ... , then you can rely on it";
- "We have only used the calorimeter for this analysis, because we are not yet confident with the central detector";
- The evening before I had to talk about this subject, I overheard the following conversation in the CERN cafeteria:
  - Young fellow: *"I have measured the resistivity, and it turns out to be* $10^{11}\,\Omega$*"*;
  - Senior: *"No, it cannot be. Tomorrow I will make the measurement and I am sure to get the right value. ... By the way, have you considered that ... ?"*

The role of beliefs in physics has been highlighted out in a particularly efficient way by the science historian Peter Galison [37]:

> *"Experiments begin and end in a matrix of beliefs. ... beliefs in instrument type, in programs of experiment enquiry, in the trained, individual judgments about every local behaviour of pieces of apparatus."*

Then, taking as an example the discovery of the positron, he remarks:

> *"Taken out of time there is no sense to the judgment that Anderson's track 75 is a positive electron; its textbook reproduction has been denuded of the prior experience that made Anderson confident in the cloud chamber, the magnet, the optics, and the photography."*

This means that pure observation does not create, or increase, knowledge without personal inputs which are needed to elaborate the information.[8] In fact, there is nothing really objective in physics, if by objective we mean that something follows necessarily from observation, like the proof of a theorem. There are, instead, beliefs everywhere. Nevertheless, physics is objective, or at least that part of it that is at present well established, if we mean by 'objective', that a rational individual cannot avoid believing it. This is the reason why we can talk in a relaxed way about beliefs in physics without even remotely thinking that it is at the same level as the stock exchange, betting on football scores, or ... New Age. The reason is that, after centuries of experimentation, theoretical work and successful predictions, there is such a consistent network of beliefs, it has acquired the status of an objective construction: one cannot mistrust one of the elements of the network without contradicting many others. Around this solid core of objective knowledge there are fuzzy borders which correspond to areas of present investigations, where the level of intersubjectivity is still very low. Nevertheless, when one proposes a new theory or model, one has to check immediately whether it contradicts some well-established beliefs. An interesting example comes from the 1997 HERA high Q^2 events, already discussed in Section 1.9. A positive consequence of this claim was to trigger a kind of mega-exercise undertaken by many theorists, consisting of systematic cross-checks of HERA data, candidate theories, and previous experimental data. The conclusion is that the most influential physicists[9] tend not to

[8] Recently, I met an elderly physicist at the meeting of the Italian Physical Society, who was nostalgic about the good old times when we could see $\pi \to \mu \to$ e decay in emulsions, and complained that at present the sophisticated electronic experiments are based on models. It took me a while to convince him that in emulsions as well he had a model and that he was not seeing these particles either.

[9] Outstanding physicists have no reluctance in talking explicitly about beliefs. Then, paradoxically, objective science is for those who avoid the word 'belief' nothing but the set of beliefs of the influential scientists to which they believe ...

believe a possible explanation in terms of new physics [64, 65]. But this has little to do with the statistical significance of the events. It is more a question of the difficulty of inserting this evidence into what is considered to be the most likely network of beliefs.

I would like to conclude this section with a Feynman quotation [66].

> *"Some years ago I had a conversation with a layman about flying saucers - because I am scientific I know all about flying saucers! I said 'I don't think there are flying saucers'. So my antagonist said, 'Is it impossible that there are flying saucers? Can you prove that it's impossible?' 'No', I said, 'I can't prove it's impossible. It's just very unlikely'. At that he said, 'You are very unscientific. If you can't prove it impossible then how can you say that it's unlikely?' But that is the way that is scientific. It is scientific only to say what is more likely and what less likely, and not to be proving all the time the possible and impossible. To define what I mean, I might have said to him, 'Listen, I mean that from my knowledge of the world that I see around me, I think that it is much more likely that the reports of flying saucers are the results of the known irrational characteristics of terrestrial intelligence than of the unknown rational efforts of extra-terrestrial intelligence'. It is just more likely. That is all."*

8.5 Biased Bayesian estimators and Monte Carlo checks of Bayesian procedures

This problem has already been raised in Sections 5.2.2 and 5.2.3. We have seen there that the expected value of a parameter can be considered, somehow, to be analogous to the estimators[10] of the frequentistic approach. It is well known, from courses on conventional statistics, that one of the nice properties an estimator should have is that of being free of bias.

Let us consider the case of Poisson and binomial distributed observations, exactly as they have been treated in Sections 5.5.1 and 5.5.2, i.e. assuming a uniform prior. Using the typical notation of frequentistic analysis, let us indicate with θ the parameter to be inferred, with $\hat{\theta}$ its estimator.

Poisson: $\theta = \lambda$; X indicates the possible observation and $\hat{\theta}$ is the estimator in the light of X:

$$\begin{aligned} \hat{\theta} &= \mathrm{E}[\lambda \,|\, X] = X + 1\,, \\ \mathrm{E}[\hat{\theta}] &= \mathrm{E}[X+1] = \lambda + 1 \neq \lambda\,. \end{aligned} \tag{8.3}$$

The estimator is biased, but consistent (the bias become negligible when X is large).

Binomial: $\theta = p$; after n trials one may observe X favourable results, and the estimator of p is then

$$\begin{aligned} \hat{\theta} &= \mathrm{E}[p \,|\, X] = \frac{X+1}{n+2}\,, \\ \mathrm{E}[\hat{\theta}] &= \mathrm{E}\left[\frac{X+1}{n+2}\right] = \frac{n\,p+1}{n+2} \neq p\,. \end{aligned} \tag{8.4}$$

In this case as well the estimator is biased, but consistent.

[10] It is worth remembering that, in the Bayesian approach, the complete answer is given by the final distribution. The prevision ('expected value') is just a way of summarizing the result, together with the standard uncertainty. Besides motivations based on penalty rules, which we cannot discuss, a practical justification is that what matters for any further approximated analysis, are expected values and standard deviation, whose properties are used in uncertainty propagation. There is nothing wrong in providing the mode(s) of the distribution or any other quantity one finds it sensible to summarize $f(\mu)$ as well.

What does it mean? The result looks worrying at first sight, but, in reality, it is the analysis of bias that is misleading. In fact:

- the initial intent is to reconstruct at best the parameter, i.e. the true value of the physical quantity identified with it;
- the freedom from bias requires only that the expected value of the estimator should equal the value of the parameter, for a given value of the parameter,

$$\begin{aligned} \mathrm{E}[\hat{\theta}\,|\,\theta] &= \theta \qquad (\text{e.g. } \mathrm{E}[\hat{\lambda}\,|\,\lambda] = \lambda)\,, \\ (\text{i.e. } \int \hat{\theta}\, f(\hat{\theta}\,|\,\theta)\, d\hat{\theta} &= \theta\ )\,. \end{aligned} \tag{8.5}$$

But what is the true value of θ? We don't know, otherwise we would not be wasting our time trying to estimate it (always keep real situations in mind!). For this reason, our considerations cannot depend only on the fluctuations of $\hat{\theta}$ around θ, but also on the different degrees of belief of the possible values of θ. Therefore they must depend also on $f_\circ(\theta)$. For this reason, the Bayesian result is that which makes the best use[11] of the state of knowledge about θ and of the distribution of $\hat{\theta}$ for each possible value θ. This can be easily understood by going back to the examples of Section 1.7. It is also easy to see that the freedom from bias of the frequentistic approach requires $f_\circ(\theta)$ to be uniformly distributed from $-\infty$ to $+\infty$ (implicitly, as frequentists refuse the very concept of probability of θ). Essentially, whenever a parameter has a limited range, the frequentistic analysis decrees that Bayesian estimators are biased.

There is another important and subtle point related to this problem, namely that of the Monte Carlo check of Bayesian methods. Let us consider the case depicted in Fig. 1.3 and imagine making a simulation, choosing the value $\mu_\circ = 1.1$, generating many (e.g. 10 000) events, and considering three different analyses:

1. a maximum likelihood analysis;
2. a Bayesian analysis, using a flat distribution for μ;
3. a Bayesian analysis, using a distribution of μ 'of the kind' $f_\circ(\mu)$ of Fig. 1.3, assuming that we have a good idea of the kind of physics we are doing.

Which analysis will reconstruct a value closest to $\mu_\circ$? You don't really need to run the Monte Carlo to realize that the first two procedures will perform equally well, while the third one, advertised as the best in these notes, will systematically underestimate $\mu_\circ$!

Now, let us assume we have observed a value of x, for example $x = 1.1$. Which analysis would you use to infer the value of μ? Considering only the results of the Monte Carlo simulation it seems obvious that one should choose one of the first two, but certainly not the third!

This way of thinking is wrong, but unfortunately it is often used by practitioners who have no time to understand what is behind Bayesian reasoning, who perform some Monte Carlo tests, and decide that the Bayesian theorem does not work![12] The solution to this apparent paradox is simple. If you believe that μ is distributed like $f_\circ(\mu)$ of Fig. 1.3, then you should use this

[11] I refer to the steps followed in the proof of Bayes' theorem given in Section 2.7. They should convince the reader that $f(\theta\,|\,\hat{\theta})$ calculated in this way is the best we can say about θ. Some say that in the Bayesian inference the answer is the answer (I have heard this sentence from A. Smith at the Valencia-6 conference), in the sense that one can use all his best knowledge to evaluate the probability of an event, but then, whatever happens, cannot change the assessed probability, but, at most, it can — and must — be taken into account for the next assessment of a different, although analogous event.

[12] This is an actual statement I have heard by Monte Carlo-oriented HEP yuppies.

distribution in the analysis and also in the generator. Making a simulation based only on a single true value, or on a set of points with equal weight, is equivalent to assuming a flat distribution for μ and, therefore, it is not surprising that the most grounded Bayesian analysis is that which performs worst in the simple-minded frequentistic checks. It is also worth remembering that priors are not just mathematical objects to be plugged into Bayes' theorem, but must reflect prior knowledge. Any inconsistent use of them leads to paradoxical results.

8.6 Frequentistic coverage

Another prejudice toward Bayesian inference shared by practitioners who have grown up with conventional statistics is related to the so-called 'frequentistic coverage'. Since, in my opinion, this is a kind of condensate of frequentistic nonsense,[13] I avoid summarizing it in my own words, as the risk of distorting something in which I cannot see any meaning is too high. A quotation[14] taken from Ref. [68] should clarify the issue:

> *"Although particle physicists may use the words 'confidence interval' loosely, the most common meaning is still in terms of original classical concept of "coverage" which follows from the method of construction suggested in Fig. ... This concept is usually stated (too narrowly, as noted below) in terms of a hypothetical ensemble of similar experiments, each of which measures m and computes a confidence interval for m_t with say, 68% C.L. Then the classical construction guarantees that in the limit of a large ensemble, 68% of the confidence intervals contain the unknown true value m_t, i.e., they 'cover' m_t. This property, called coverage in the frequentistic sense, is the defining property of classical confidence intervals. It is important to see this property as what it is: it reflects the relative frequency with which the statement, 'm_t is in the interval (m_1, m_2)', is a true statement. The probabilistic variables in this statements are m_1 and m_2; m_t is fixed and unknown. It is equally important to see what frequentistic coverage is not: it is a not statement about the degree of belief that m_t lies within the confidence interval of a particular experiment. The whole concept of 'degree of belief' does not exist with respect to classical confidence intervals, which are cleverly (some would say devilishly) defined by a construction which keeps strictly to statements about $P(m \mid m_t)$ and never uses a probability density in the variable m_t.*
>
> *This strict classical approach can be considered to be either a virtue or a flaw, but I think that both critics and adherents commonly make a mistake in describing coverage from the narrow point of view which I described in the preceeding paragraph. As Neyman himself pointed out from the beginning, the concept of coverage is not restricted to the idea of an ensemble of hypothetical nearly-identical experiments. Classical confidence intervals have a much more powerful property: if, in an ensemble of real, different, experiments, each experiment measures whatever observables it likes, and construct a 68% C.L. confidence interval, then in the long run 68% of the confidence intervals cover the true value of their respective observables. This is directly applicable to real life, and is the real beauty of classical confidence intervals."*

I think that the reader can judge for himself whether this approach seems reasonable. From the Bayesian point of view, the full answer is provided by $P(m_t \mid m)$, to use the same notation of Ref. [68]. If this evaluation has been carried out under the requirement of coherence, from $P(m_t \mid m)$ one can evaluate a probability for m_t to lie in the interval (m_1, m_2). If this probability is 68%, in order to stick to the same value this implies:

[13] Zech says, more optimistically: *"Coverage is the magic objective of classical confidence bounds. It is an attractive property from a purely esthetic point of view but it is not obvious how to make use of this concept."*[67]

[14] The translation of the symbols is as follows: m stands for the measured quantity (x or $\hat{\theta}$ in these notes); m_t stands for the true value (μ or θ here); $P(\cdot \mid \cdot)$ for $f(\cdot \mid \cdot)$.

- one believes 68% that m_t is in that interval;
- one is ready to place a $\approx 2:1$ bet on m_t being in that interval and a $\approx 1:2$ bet on m_t being elsewhere;
- if one imagines n situations in which one has similar conditions (they could be different experiments, or simply urns containing a 68% proportion of white balls) and thinks of the relative frequency with which one expects that this statement will be true (f_n), logic applied to the basic rules of probability imply that, with the increasing n, it will become more and more improbable that f_n will differ much from 68% (Bernoulli theorem).

So, the intuitive concept of 'coverage' is naturally included in the Bayesian result and it is expressed in intuitive terms (probability of true value and expected frequency). But this result has to depend also on priors, as seen in the previous section and in many other places in this report (see, for example, Section 1.7). Talking about coverage independently of prior knowledge (as frequentists do) makes no sense, and leads to contradictions and paradoxes. Imagine, for example, an experiment operated for one hour at LEP200 and reporting zero candidate events for zirconium production in e^+e^- in the absence of expected background. I do not think that there is a single particle physicist ready to believe that, if the experiment is repeated many times, in only 68% of the cases the 68% C.L. interval [0.00, 1.29] will contain the true value of the 'Poisson signal mean', as a blind use of Table II of Ref. [60] would imply.[15] If this example seems a bit odd, I invite you to think about the many 95% C.L. lower limits on the mass of postulated particles. Do you really believe that in 95% of the cases the mass is above the limit, and in 5% of the cases below the limit? If this is the case, you would bet \$5 on a mass value below the limit, and receive \$100 if this happened to be true (you should be ready to accept the bet, since, if you believe in frequentistic coverage, you must admit that the bet is fair). But perhaps you will never accept such a bet because you believe much more than 95% that the mass is above the limit, and then the bet is not fair at all; or because you are aware of thousands of lower limits, and a particle has never shown up on the 5% side ...

8.7 Bayesian networks

In Section 8.4 I mentioned the network of beliefs which give the perceived status of objectivity to consolidated science. In fact, belief networks, also called Bayesian networks, are not only an abstract idea useful in epistemology. They represent one of the most promising applications of Bayesian inference and they have generated a renewed interest in the field of artificial intelligence, where they are used for expert systems, decision makers, etc. [69].

Although, to my knowledge, there are not yet specific HEP applications of these methods, I would like to give a rough idea of what they are and how they work, with the help of a simple example. You are visiting some friends, and, minutes after being in their house, you sneeze. You know you are allergic to pollen and to cats, but it could also be a cold. What is the cause of the sneeze? Figure 8.1 sketches the problem. There are some facts about which you are sure (the sneeze, the weather conditions and the season), but you don't know if the sneeze is a symptom of a cold or of an allergy. In particular, you don't know if there is a cat in the house.

[15] One would object that this is, more or less, the result that we could obtain making a Bayesian analysis with a uniform prior. But it was said that this prior assumes a positive attitude of the experimenters, i.e. that the experiment was planned, financed, and operated by rational people, with the hope of observing something (see Sections 5.4.3 and 5.5.2). This topic, together with the issue of reporting experimental results in a prior-free way, is discussed in detail in Ref. [25].

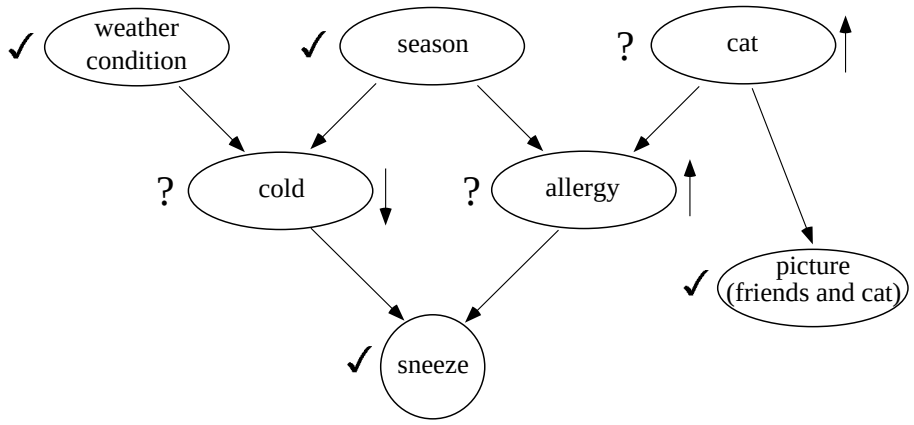

Figure 8.1: An example of belief network.

Then, you see a picture of your friend with a cat. This could be an indication that they have a cat, but it is just an indication. Nevertheless, this indication increases the probability that there is a cat around, and then the probability that the cause of the sneeze is cat's hair allergy increases, while the probability of any other potential cause decreases. If you then establish with certainty the presence of the cat, the cause of the allergy also becomes practically certain.

The idea of Bayesian networks is to build a network of causes and effects. Each event, generally speaking, can be certain or uncertain. When there is a new piece of evidence, this is transmitted to the whole network and all the beliefs are updated. The research activity in this field consists of the most efficient way of doing the calculation, using Bayesian inference, graph theory, and numerical approximations.

If one compares Bayesian networks with other ways of pursuing artificial intelligence their superiority is rather clear: they are close to the natural way of human reasoning, the initial beliefs can be those of experts (avoiding the long training needed to set up, for example, neural networks, unfeasible in practical applications), and they learn by experience as soon as they start to receive evidence [70].

8.8 Why do frequentistic hypothesis tests 'often work'?

The problem of classifying hypotheses according to their credibility is natural in the Bayesian framework. Let us recall briefly the following way of drawing conclusions about two hypotheses in the light of some data:

$$\frac{P(H_i \,|\, \text{Data})}{P(H_j \,|\, \text{Data})} \;=\; \frac{P(\text{Data} \,|\, H_i)}{P(\text{Data} \,|\, H_j)} \cdot \frac{P_\circ(H_i)}{P_\circ(H_j)} \,. \tag{8.6}$$

This form is very convenient, because:

- it is valid even if the hypotheses H_i do not form a complete class [a necessary condition if, instead, one wants to give the result in the standard form of Bayes' theorem given by formula (3.11)];
- it shows that the Bayes factor is an unbiased way of reporting the result (especially if a different initial probability could substantially change the conclusions);

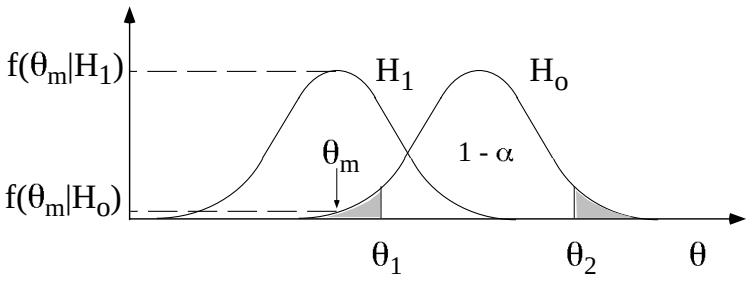

Figure 8.2: Testing a hypothesis $H_\circ$ implies that one is ready to replace it with an alternative hypothesis.

- the Bayes factor depends only on the likelihoods of observed data and not at all on unobserved data (contrary to what happens in conventional statistics, where conclusions depend on the probability of all the configurations of data in the tails of the distribution[16]). In other words, Bayes' theorem applies in the form (8.6) and not as

$$\underbrace{\frac{P(H_i\,|\,\text{Data+Tail})}{P(H_j\,|\,\text{Data+Tail})} = \frac{P(\text{Data+Tail}\,|\,H_i)}{P(\text{Data+Tail}\,|\,H_j)} \cdot \frac{P_\circ(H_i)}{P_\circ(H_j)}}_{?};$$

- testing a single hypothesis does not make sense: one may talk of the probability of the Standard Model (SM) only if one is considering an Alternative Model (AM), thus getting, for example,

$$\frac{P(\text{AM}\,|\,\text{Data})}{P(\text{SM}\,|\,\text{Data})} = \frac{P(\text{Data}\,|\,\text{AM})}{P(\text{Data}\,|\,\text{SM})} \cdot \frac{P_\circ(\text{AM})}{P_\circ(\text{SM})} :$$

  $P(\text{Data}\,|\,\text{SM})$ can be arbitrarily small, but if there is not a reasonable alternative one has only to accept the fact that some events have been observed which are very far from the expectation value;

- repeating what has been said several times, in the Bayesian scheme the conclusions depend only on observed data and on previous knowledge; in particular, they do not depend on
  - how the data have been combined;
  - data not observed and considered to be even rarer than the observed data;
  - what the experimenter was planning to do before starting to take data. (I am referring to predefined fiducial cuts and the stopping rule, which, according to the frequentistic scheme should be defined in the test protocol. Unfortunately I cannot discuss this matter here in detail and I recommend the reading of Ref. [10]).

At this point we can finally reply to the question: "why do commonly-used methods of hypothesis testing usually work?" (see Sections 1.8 and 1.9).

By reference to Fig. 8.2 (imagine for a moment the figure without the curve H_1), the argument that θ_m provides evidence against $H_\circ$ is intuitively accepted and often works, not (only) because of probabilistic considerations of θ in the light of $H_\circ$, but because it is often reasonable to imagine an alternative hypothesis H_1 that

[16] The necessity of using integrated distributions is due to the fact that the probability of observing a particular configuration is always very small, and a frequentistic test would reject the null hypotheses.

Figure 8.3: Experimental obituary (courtesy of Alvaro de Rujula [71]).

1. maximizes the likelihood $f(\theta_m \,|\, H_1)$ or, at least

$$\frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \gg 1\,;$$

2. has a comparable prior $[P_\circ(H_1) \approx P_\circ(H_\circ)]$, such that

$$\frac{P(H_1 \,|\, \theta_m)}{P(H_\circ \,|\, \theta_m)} = \frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \cdot \frac{P_\circ(H_1)}{P_\circ(H_\circ)} \quad \approx \quad \frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \longrightarrow \gg 1\,.$$

So, even though there is no objective or logical reason why the frequentistic scheme should work, the reason why it often does is that in many cases the test is made when one has serious doubts about the null hypothesis. But a peak appearing in the middle of a distribution, or any excess of events, is not, in itself, a hint of new physics (Fig. 8.3 is an invitation to meditation ...). My recommendations are therefore the following.

- Be very careful when drawing conclusions from χ^2 tests, '$3\,\sigma$ golden rule', and other 'bits of magic';
- Do not pay too much attention to fixed rules suggested by statistics 'experts', supervisors, and even Nobel laureates, taking also into account that
  - they usually have permanent positions and risk less than PhD students and postdocs who do most of the real work;
  - they have been 'miseducated' by the exciting experience of the glorious 1950s to 1970s: as Giorgio Salvini says, *"when I was young, and it was possible to go to sleep at night after having added within the day some important brick to the building of the*

elementary particle palace. We were certainly lucky." [72]. Especially when they were hunting for resonances, priors were very high, and the 3–4 σ rule was a good guide.

- Fluctuations exist. There are millions of frequentistic tests made every year in the world. And there is no probability theorem ensuring that the most extreme fluctuations occur to a precise Chinese student, rather than to a large HEP collaboration (this is the same reasoning of many Italians who buy national *lotteria* tickets in Rome or in motorway restaurants, because 'these tickets win more often' ...).

As a conclusion to these remarks, and to invite the reader to take with much care the assumption of equiprobability of hypothesis (a hidden assumption in many frequentistic methods), I would like to add this quotation by Poincaré [6]:

> *"To make my meaning clearer, I go back to the game of écarté mentioned before.*[17] *My adversary deals for the first time and turns up a king. What is the probability that he is a sharper? The formulae ordinarily taught give 8/9, a result which is obviously rather surprising. If we look at it closer, we see that the conclusion is arrived at as if, before sitting down at the table, I had considered that there was one chance in two that my adversary was not honest. An absurd hypothesis, because in that case I should certainly not have played with him; and this explains the absurdity of the conclusion. The function on the à priori probability was unjustified, and that is why the conclusion of the à posteriori probability led me into an inadmissible result. The importance of this preliminary convention is obvious. I shall even add that if none were made, the problem of the à posteriori probability would have no meaning. It must be always made either explicitly or tacitly."*

8.9 Frequentists and Bayesian 'sects'

Many readers may be interested in how the problem 'to Bayes or not to Bayes' is viewed by statisticians. In order to thoroughly analyse the situation, one should make a detailed study not only of the probability theory, but also of the history and sociology of statistical science. The most I can do here is to give personal impressions, certainly biased, and some references. I invite the reader to visit the statistics department in his University, browse their journals and books, and talk to people (and to judge the different theses by the logical strength of their arguments, not weighing them just by numbers ...).

8.9.1 Bayesian versus frequentistic methods

An often cited paper for a reasonably balanced discussion [46] on the subject is the article *"Why isn't everyone a Bayesian?"*, by B. Efron [73]. Key words of the paper are: *Fisherian inference*; *Frequentistic theory*; *Neyman–Pearson–Wald*; *Objectivity*. For this reason, pointing out this paper as 'balanced' is not really fair. Nevertheless, I recommend reading the article, together with the accompanying comments and the reply by the author published in the same issue of the journal (a typical practice amongst statisticians).

So, it is true that *"Fisherian and Neyman–Pearson–Wald ideas have shouldered Bayesian theory aside in statistical practice"* [73], but *"The answer is simply that statisticians do not know what the statistical paradigm says. Why should they? There are very few universities in the world with statistics departments that provides a good course on the subject."* [74] Essentially, the main point of the Efron paper is to maintain traditional methods, despite the *"disturbing*

[17] See Section 1.6.

catalog of inconsistencies" [73], and the *"powerful theoretical reasons for preferring Bayesian inference"* [73]. Moreover, perhaps not everybody who cites the Efron paper is aware of further discussions about it, like the letter in which Zellner [75] points out that one of the problems posed by Efron already had a Bayesian solution (in the Jeffreys' book [29]), that Efron admitted to knowing and even to having used [76]. As a kind of final comment on this debated paper, I would like to cite Efron's last published reply I am aware of [76]:

> *"First of all let me thank the writers for taking my article in its intended spirit: not as an attack on the Bayesian enterprise, but rather as a critique of its preoccupation with philosophical questions, to the detriment of statistical practice. Meanwhile I have received some papers, in particular one from A.F.M. Smith, which show a healthy Bayesian interest in applications, so my worries were overstated if not completely groundless."*

There are some other references which I would like to suggest if you are interested in forming your own opinion on the subject. They have also appeared in The American Statistician, where in 1997 an entire Teaching Corner section of the journal [63] was devoted to three papers presented in a round table on 'Bayesian possibilities for introductory statistics' at the 156th Annual Meeting of the American Statistical Association, held in Chicago, in August 1996. For me these articles are particularly important because I was by chance in the audience of the round table (really 'by chance'!). At the end of the presentations I was finally convinced that frequentism was dead, at least as a philosophical idea. I must say, I was persuaded by the non-arguments of the defender of frequentism even more than by the arguments of the defenders of the Bayesian approach. I report here the abstract[18] of Moore, who presented the 'reason to hesitate' to teach Bayesian statistics:

> *"The thesis of this paper is that Bayesian inference, important though it is for statisticians, is among the mainly important statistical topics that it is wise to avoid in most introductory instruction. The first reason is pragmatic (and empirical): Bayesian methods are as yet relatively little used in practice. We have an obligation to prepare students to understand the statistics they will meet in their further studies and work, not the statistics we may hope will someday replace now-standard methods. A second argument also reflects current conditions: Bayesians do not agree on standard approaches to standard problem settings. Finally, the reasoning of Bayesian inference, depending as it does on ideas of conditional probability, is quite difficult for beginners to appreciate. There is of course no easy path to a conceptual grasp of inference, but standard inference at least rests on repetition of one straightforward question, What would happen if I did this many times? "*

Even if some arguments might be valid, thinking about statisticians who make surveys in a standardized form (in fields that they rarely understand, such as medicine and agriculture), surely they do not hold in physics, even less in frontier physics. As I commented to Moore after his talk, what is important for a physicist is not "what would happen if I did this many times?", but "what am I learning by the experiment?".[19]

8.9.2 Orthodox teacher versus sharp student - a dialogue by Gabor

As a last comment about frequentistic ideas I would like to add here a nice dialogue, which was circulated via internet on 19th February 1999, with an introduction and comment by the

[18] I quote here the original abstract, which appears on page 18 of the conference abstract book.

[19] I also made other comments on the general illogicality of his arguments, which you may easily imagine by reading the abstract. For these comments I even received applause from the audience, which really surprised me, until I learned that David Moore is one of the most authoritative American statisticians: only a outsider like me would have said what I said ...

author, the statistician George Gabor [77] of Dalhousie University (Halifax, N.S., Canada). It was meant as a contribution to a discussion triggered by D.A. Berry (that of Refs. [10] and [63]) a few days before.

> *"Perhaps a Socratic exchange between an ideally sharp, i.e not easily bamboozled student (S.) of a typical introductory statistics course and his prof (P.) is the best way to illustrate what I think of the issue. The class is at the point where confidence interval (CI) for the normal mean is introduced and illustrated with a concrete example for the first time.*
>
> **P.** *...and so a 95% CI for the unknown mean is (1.2, 2.3).*
>
> **S.** *Excuse me sir, just a few minutes ago you emphasized that a CI is some kind of random interval with certain coverage properties in REPEATED trials.*
>
> **P.** *Correct.*
>
> **S.** *What, then, is the meaning of the interval above?*
>
> **P.** *Well, it is one of the many possible realizations from a collection of intervals of a certain kind.*
>
> **S.** *And can we say that the 95 collective, is somehow carried over to this particular realization?*
>
> **P.** *No, we can't. It would be worse than incorrect; it would be meaningless for the probability claim is tied to the collective.*
>
> **S.** *Your claim is then meaningless?*
>
> **P.** *No, it isn't. There is actually a way, called Bayesian statistics, to attribute a single-trial meaning to it, but that is beyond the scope of this course. However, I can assure you that there is no numerical difference between the two approaches.*
>
> **S.** *Do you mean they always agree?*
>
> **P.** *No, but in this case they do provided that you have no reason, prior to obtaining the data, to believe that the unknown mean is in any particularly narrow area.*
>
> **S.** *Fair enough. I also noticed sir that you called it 'a' CI, instead of 'the' CI. Are there others then?*
>
> **P.** *Yes, there are actually infinitely many ways to obtain CI's which all have the same coverage properties. But only the one above is a Bayesian interval (with the proviso above added, of course).*
>
> **S.** *Is Bayesian-ness the only way to justify the use of this particular one?*
>
> **P.** *No, there are other ways too, but they are complicated and they operate with concepts that draw their meaning from the collective (except the so called likelihood interval, but then this strange guy does not operate with probability at all).*
>
> **...**
>
> *It could be continued ad infinitum. Assuming sufficiently more advanced students one could come up with similar exchanges concerning practically every frequentist concept orthodoxy operates with (sampling distribution of estimates, measures of performance, the very concept of independence, etc.). The point is that orthodoxy would fail at the first opportunity had students been sufficiently sharp, open minded, and inquisitive. That we are not humiliated repeatedly by such exchanges (in my long experience not a single one has ever taken place) says more about... well, I don't quite know about what — the way the mind plays tricks with the concept of probability? The background of our students? Both?*
>
> *Ultimately then we teach the orthodoxy not only because of intellectual inertia, tradition, and the rest; but also because, like good con artists, we can get away with it. And that I find very disturbing. I must agree with Basu's dictum that nothing in orthodox statistics makes sense unless it has a Bayesian interpretation. If, as is the case, the only thing one*

can say about frequentist methods is that they work only in so far as they don't violate the likelihood principle; and if they don't (and they frequently do), they numerically agree with a Bayesian procedure with some flat prior - then we should go ahead and teach the real thing, not the substitute. (The latter, incidentally, can live only parasitically on an illicit Bayesian usage of its terms. Just ask an unsuspecting biologist how he thinks about a CI or a P-value.)

One can understand, or perhaps follow is a better word, the historical reasons orthodoxy has become the prevailing view. Now, however, we know better."

8.9.3 Subjective or objective Bayesian theory?

Once you have understood that probability and frequencies are different concepts, that probability of hypothesis is a useful and natural concept for reporting results, that Bayes' theorem is a powerful tool for updating probability and learning from data, that priors are important and pretending that they do not exist is equivalent to assuming them flat, and so on, it is difficult to then take a step back. However, it is true that there is no single shared point of view among those who, generally speaking, support the Bayesian approach. I don't pretend that I can provide an exhaustive analyse of the situation here, or to be unbiased about this matter either.

The main schools of thought are the 'subjectivists' and the 'objectivists'. The dispute may look strange to an outsider, if one thinks that both schools use probability to represent degrees of belief. Nevertheless, objectivists want to minimize the person's contribution to the inference, by introducing reference priors (for example Jeffreys' priors [29]) or other constraints, such as maximum entropy (for an overview see Refs. [19] and [78]). The motto is *"let the data speak for themselves"*. I find this subject highly confusing, and even Bernardo and Smith (Bernardo is one of the key persons behind reference priors) give the impression of contradicting themselves often on this point as, for example, when the subject of reference analysis is introduced:

"to many attracted to the formalism of the Bayesian inferential paradigm, the idea of a non-informative prior distribution, representing 'ignorance' and 'letting the data speak for themselves' has proved extremely seductive, often being regarded as synonymous with providing objective inferences. It will be clear from the general subjective perspective we have maintained throughout this volume, that we regard this search for 'objectivity' to be misguided. However, it will also be clear from our detailed development in Section 5.4 that we recognize the rather special nature and role of the concept of a 'minimal informative' prior specification - appropriately defined! In any case, the considerable body of conceptual and theoretical literature devoted to identifying 'appropriate' procedures for formulating prior representations of 'ignorance' constitutes a fascinating chapter in the history of Bayesian Statistics. In this section we shall provide an overview of some of the main directions followed in this search for a Bayesian 'Holy Grail'.[19]

In my point of view, the extreme idea along this line is represented by the Jaynes' 'robot' (*"In order to direct attention to constructive things and away from controversial irrelevance, we shall invent an imaginary being. Its brain is to be designed by us, so that it reasons according to certain defined rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains"* [79]).

As far as I understand it, I see only problems with objectivism, although I do agree on the notion of a commonly perceived objectivity, in the sense of intersubjectivity (see Section 8.4). Frankly, I find probabilistic evaluations made by a coherent subjectivist, assessed under personal

responsibility, to be more trustworthy and more objective than values obtained in a mechanical way using objective prescriptions [22].

Moving to a philosophical level deeper than this kind of angels' sex debate (see Section 3.6), there is the important issue of what an event is. All events listed in Section 8.1 (apart from that of point 4) are somehow verifiable. Perhaps one will have to wait until tomorrow, the end of 1999, or 2010, but at a certain point the event may become certain, either true or false. However, one can think about other events, examples of which have been shown in these notes, that are not verifiable, either for a question of principle, or by accident.

- The old friend could die, carrying with him the secret of whether he had been cheating, or simply lucky (Section 3.4.5).

- The particle interacts with the detector (Section 3.4.4) and continues its flight: was it really a π or a μ?

- Using our best knowledge about temperature measurement we can state that the temperature of a room at a certain instant is $21.7 \pm 0.3\,^{\circ}\mathrm{C}$ with 95% probability (Section 8.1); after the measurement the window is opened, the weather changes, the thermometer is lost: how is it possible to verify the event '$21.4 \leq T/^{\circ}\mathrm{C} \leq 22.0$'?

This problem is present every time we make a probabilistic statement about physics quantities. It is present not only when a measurand is critically time dependent (the position of a plane above the Atlantic), but also in the case of fundamental constants. In this latter case we usually believe in the progress of science and thus we hope that the quantity will be measured so well in the future that it will one day become a kind of exact value, in comparison to today's uncertainty. But it is absurd to think that one day we will be able to 'open an electron' and read on a label all its properties with an infinite number of digits. This means that for scientific applications it is convenient to enlarge the concept of an event (see Section 3.3.2), releasing the condition of verifiability.[20] At this point the normative role of the hypothetical coherent bet becomes crucial. A probability evaluation, made by an honest person well-trained in applying coherence on verifiable events, becomes, in my opinion, the only means by which degrees of belief can be exchanged among rational people. We have certainly reached a point in which the domain of physics, metaphysics and moral overlap, but it looks to me that this is exactly the way in which science advances.

It seems to me that almost all Bayesian schools support this idea of the extended meaning of an event, explicitly or tacitly (anyone who speaks about $f(\theta)$, with θ a parameter of a distribution, does it). A more radical point of view, which is very appealing from the philosophical perspective, but more difficult to apply (at least in physics), is the predictive approach (or operational subjectivism), along the lines of de Finetti's thinking. The concept of probability is strictly applied only to real observables, very precisely ('operationally') defined. The events are all associated with discrete uncertain numbers (integer or rational), in the simplest case 1 or 0 if there are only two possibilities (true or false). Having excluded non-observables, it makes no sense to speak of $f(\mu \,|\, \mathrm{data})$, but only of $f(x \,|\, \mathrm{data})$, where X stands for a future (or, in general, not yet known) observation. For the moment I prefer to stick to our 'metaphysical' true values, but I encourage anyone who is interested in this subject to read Lad's recent book [80], which also contains a very interesting philosophical and historical introduction to the subject.

[20] It is interesting to realize, in the light of this reflection, that the ISO definition of true value (*"a value compatible with the definition of a given particular quantity"*, see Sections 1.2 and 1.3) can accommodate this point of view.

8.9.4 Bayes' theorem is not all

Finally, I would like to recall that Bayes' theorem is a very important tool, but it can be used only when the scheme of prior, likelihood, and final is set up, and the distributions are properly normalized.[21] This happens very often in measurement uncertainty problems, but less frequently in other applications, such as assessing the probabilities of hypotheses. When Bayes' theorem is not applicable, conclusions may become strongly dependent on individuals and the only guidance remains the normative rule of the hypothetical coherent bet.

8.10 Solution to some problems

Here are the solutions to some of the examples of the notes.

8.10.1 AIDS test

The AIDS test problem (Example 7 of Section 1.9) is a very standard one. Let us solve it using the Bayes factor:

$$\begin{aligned}
\frac{P(\mathrm{HIV}\,|\,\mathrm{Positive})}{P(\overline{\mathrm{HIV}}\,|\,\mathrm{Positive})} &= \frac{P(\mathrm{Positive}\,|\,\mathrm{HIV})}{P(\mathrm{Positive}\,|\,\overline{\mathrm{HIV}})}\cdot\frac{P_\circ(\mathrm{HIV})}{P(\overline{\mathrm{HIV}})} \\
&= \frac{\approx 1}{0.002}\times\frac{0.1/60}{\approx 1} = 500\times\frac{1}{600} = \frac{1}{1.2} \\
P(\mathrm{HIV}\,|\,\mathrm{Positive}) &= 45.5\%\,.
\end{aligned}$$

Writing Bayes' theorem in this way helps a lot in understanding what is going on. Stated in terms of signal to noise and selectivity (see problem 1 in Section 3.4.4), we are in a situation in which the selectivity of the test is not enough for the noisy conditions. So in order to be practically sure that the patient declared 'positive' is infected, with this performance of the analysis, one needs independent tests, unless the patient belongs to high-risk classes. For example, a double independent analysis on an average person would yield

$$P(\mathrm{HIV}\,|\,\mathrm{Positive}_1\cap\mathrm{Positive}_2) = 99.76\%\,,$$

similar[22] to that obtained in the case where a physician had a 'severe doubt' (i.e. $P_\circ(\mathrm{HIV})\approx P_\circ(\overline{\mathrm{HIV}})$ that the patient could be infected:

$$P(\mathrm{HIV}\,|\,\mathrm{Positive}, P_\circ(\mathrm{HIV})\approx 0.5) = 99.80\%\,.$$

We see then that, as discussed several times (see Section 8.8), the conclusion obtained by arbitrary probability inversion is equivalent to assuming uniform priors.

[21] I have made use several times in these notes of improper distributions, i.e. such that

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x \to \infty\,,$$

but, as specified, they were always thought to be the limit of proper distributions (see, for example, Section 5.5.2).

[22] There is nothing profound in the fact that the two cases give very similar results. It is just due to the numbers of these examples (i.e. $500\approx 600$).

8.10.2 Gold/silver ring problem

The three-box problem (Section 3.4.4) seems to be intuitive for some, but not for everybody. Let us label the three boxes: A, Golden-Golden; B, Golden-Silver; C, Silver-Silver. The initial probability (i.e. before having checked the first ring) of having chosen the box A, B, or C is, by symmetry, $P_\circ(A) = P_\circ(B) = P_\circ(C) = 1/3$.

This probability is updated after the event $E =$'the first ring extracted is golden' by Bayes' theorem:

$$\begin{aligned}
P(A\,|\,E) &= \frac{P(E\,|\,A)\cdot P_\circ(A)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 2/3 \\
P(B\,|\,E) &= \frac{P(E\,|\,B)\cdot P_\circ(B)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 1/3 \\
P(C\,|\,E) &= \frac{P(E\,|\,C)\cdot P_\circ(C)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 0\,,
\end{aligned}$$

where $P(E\,|\,A)$, $P(E\,|\,B)$ and $P(E\,|\,C)$ are, respectively, 1, 1/2 and 0.

Finally, calling $F =$'the next ring will be golden if I extract it from the same box', we have, using the probability rules:

$$\begin{aligned}
P(F\,|\,E) &= P(F|A,E)\cdot P(A|E) + P(F|B,E)\cdot P(B|E) + P(F|C,E)\cdot P(C|E) \\
&= 1\times 2/3 + 0\times 1/3 + 0\times 0 = 2/3\,.
\end{aligned}$$

Chapter 9

Further HEP applications

9.1 Poisson model: dependence on priors, combination of results and systematic effects

The inference on the parameter λ of the Poisson has been treated in Sections 5.5.2 and 5.6.5. Here we will take a look at other applications of practical interest.

9.1.1 Dependence on priors

The results of Sections 5.5.2 and 5.6.5 were obtained using a uniform prior. One may worry how much the result changes if different priors are used in the analysis. Bearing in mind the rule of coherence, we are clearly interested only in reasonable[1] priors.

In frontier physics the choice of $f_\circ(\lambda) = k$ is often not reasonable. For example, searching for monopoles, one does not believe that $\lambda = 10^6$ and $\lambda = 1$ are equally possible. Realistically, one would expect to observe, with the planned experiment and running time, $\mathcal{O}(10)$ monopoles, if they exist at all. We follow the same arguments of Section 5.4.3 (negative neutrino mass), modelling the prior beliefs of a community of rational people who have planned and run the experiment. For reasons of mathematical convenience, we model $f_\circ(\lambda)$ with an exponential, but, extrapolating the results of Section 5.4.3, it is easy to understand that the exact function is not really crucial for the final result.

The function

$$f_\circ(\lambda) = \frac{1}{10} e^{-\lambda/10}\,, \tag{9.1}$$

with

$$\begin{aligned} \mathrm{E}_\circ[\lambda] &= 10 \\ \sigma_\circ(\lambda) &= 10 \end{aligned}$$

may be well suited to the case: the highest beliefs are for small values of λ, but also values up

[1] I insist on the fact that they must be reasonable, and not just any prior. The fact that absurd priors give absurd results does not invalidate the inferential framework based on subjective probability.

to 30 or 50 would not be really surprising. We obtain the following results:

$$\begin{aligned} f(\lambda \,|\, x=0) &= \frac{e^{-\lambda}\frac{1}{10}\,e^{-\lambda/10}}{\int_0^\infty(\dots)\mathrm{d}\lambda} && (9.2)\\ &= \frac{11}{10}\,e^{-\frac{11}{10}\lambda} && (9.3)\\ \mathrm{E}[\lambda] &= 0.91 \\ P(\lambda \le 2.7) &= 95\% \\ \lambda_u &= 2.7 \text{ with } 95\% \text{ probability}\,. && (9.4) \end{aligned}$$

The result is very stable. Changing $\mathrm{E}_\circ[\lambda]$ from '∞' to 10 has only a 10% effect on the upper limit. As far as the scientific conclusions are concerned, the two limit are identical. For this reason one should not worry about using a uniform prior, and complicate one's life to model a more realistic prior.

As an exercise, we can extend this result to a generic expected value of events, still sticking to the exponential:

$$f_\circ(\lambda) = \frac{1}{\lambda_\circ}e^{-\lambda/\lambda_\circ}\,,$$

which has an expected value $\lambda_\circ$. The uniform distribution is recovered for $\lambda_\circ \to \infty$. We get:

$$\begin{aligned} f(\lambda \,|\, x=0,\lambda_\circ) &\propto e^{-\lambda}\frac{1}{\lambda_\circ}e^{-\lambda/\lambda_\circ} \\ f(\lambda \,|\, x=0,\lambda_\circ) &= (1+\lambda_\circ)e^{-\lambda\,(1+\lambda_\circ)/\lambda_\circ} \\ &= \frac{1}{\lambda_1}e^{-\lambda/\lambda_1} \\ \text{with } \frac{1}{\lambda_1} &= \frac{1}{1}+\frac{1}{\lambda_\circ} \\ F(\lambda \,|\, x=0,\lambda_\circ) &= 1-e^{-\lambda/\lambda_\circ}\,. \end{aligned}$$

The upper limit, at a probability level P_u, becomes:

$$\lambda_u = -\lambda_1\,\ln(1-P_u)\,. \qquad (9.5)$$

9.1.2 Combination of results from similar experiments

Results may be combined in a natural way making an interactive use of Bayesian inference. As a first case we assume several experiments having the same efficiency and exposure time.

- Prior knowledge:
$$f_\circ(\lambda \,|\, I_\circ)\,;$$
- Experiment 1 provides Data_1:
$$f_1(\lambda \,|\, I_\circ, \mathrm{Data}_1) \propto f(\mathrm{Data}_1 \,|\, \lambda, I_\circ)\cdot f_\circ(\lambda \,|\, I_\circ)\,;$$
- Experiment 2 provides Data_2:
$$f_2(\lambda \,|\, I_\circ, \mathrm{Data}_1 \dots) \propto f(\mathrm{Data}_2 \,|\, \lambda, I_\circ)\cdot f_1(\lambda \,|\, \dots)\,;$$
$$\Rightarrow f_2(\lambda \,|\, I_\circ, \mathrm{Data}_1, \mathrm{Data}_2)\,.$$

- Combining n similar independent experiments we get

$$\begin{aligned} f(\lambda \,|\, \underline{x}) &\propto \Pi_{i=1}^{n} f(x_i \,|\, \lambda) \cdot f_\circ(\lambda) \\ &\propto f(\underline{x} \,|\, \lambda) \cdot f_\circ(\lambda) \\ &\propto \Pi_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \cdot f_\circ(\lambda) \\ &\propto e^{-n\,\lambda}\lambda^{\sum_{i=1}^{n} x_i} f_\circ(\lambda)\,. \end{aligned} \tag{9.6}$$

  Then it is possible to evaluate expected value, standard deviation, and probability intervals.

As an exercise, let us analyse the two extreme cases, starting from a uniform prior:

$\boxed{\sum_i x_i = 0}$ if none of the n <u>similar</u> experiments has observed events we have

$$\begin{aligned} f(\lambda \,|\, n \text{ expts}, 0 \text{ evts}) &= n e^{-n\,\lambda} \\ F(\lambda \,|\, n \text{ expts}, 0 \text{ evts}) &= 1 - e^{-n\,\lambda} \\ \lambda_u &= -\frac{\ln(1-P_u)}{n} \quad \text{with probability } P_u\,. \end{aligned}$$

$\boxed{\sum_i x_i \text{ \textbf{“large”}}}$ If the number of observed events is large (and the prior flat), the result will be normally distributed:

$$f(\lambda) \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)\,.$$

Then, in this case it is more practical to use maximum likelihood methods than to make integrals (see Section 2.9). From the maximum of $f(\lambda)$, in correspondence of $\lambda = \lambda_m$, we easily get:

$$\mu_\lambda = \mathrm{E}(\lambda) \approx \lambda_m = \frac{\sum_{i=1}^{n} x_i}{n}\,,$$

and from the second derivative of $\ln f(\lambda)$ around the maximum:

$$\begin{aligned} \left.\frac{\partial^2 \ln f(\lambda)}{\partial \lambda^2}\right|_{\lambda_m} &= \frac{-n^2}{\sum_{i=1}^{n} x_i} \\ \sigma_\lambda^2 &\approx -\left(\left.\frac{\partial^2 \ln f(\lambda)}{\partial \lambda}\right|_{\lambda_m}\right)^{-1} = \frac{1}{n}\frac{\sum_{i=1}^{n} x_i}{n} \\ \sigma_\lambda &\approx \frac{\sqrt{\mu_\lambda}}{\sqrt{n}}\,. \end{aligned}$$

9.1.3 Combination of results: general case

The previous case is rather artificial and can be used, at most, to combine several measurements of the same experiment repeated n times, each with the same running time. In general, experiments differ in size, efficiency, and running time. A result on λ is no longer meaningful. The quantity which is independent from these contingent factors is the rate, related to λ by

$$r = \frac{\lambda}{\epsilon\, S\, \Delta T} = \frac{\lambda}{\mathcal{L}}\,,$$

where ϵ indicates the efficiency, S the generic 'size' (either area or volume, depending on whatever is relevant for the kind of detection) and ΔT the running time: all the factors have been grouped into a generic 'integrated luminosity' $\mathcal{L}$ which quantify the effective exposure of the experiment.

As seen in the previous case, the combined result can be achieved using Bayes' theorem iteratively, but now one has to pay attention to the fact that:

- the observable is Poisson distributed, and the each experiment can infer a λ parameter;
- the result on λ must be translated[2] into a result on r.

Starting from a prior on r (e.g. a monopole flux) and going from experiment 1 to n we have

- from $f_\circ(r)$ and $\mathcal{L}_1$ we get $f_\circ(\lambda)$; then, from the data we perform the inference on λ and then on r:

$$\begin{aligned} f_\circ(r)\,\&\,\mathcal{L}_1 &\rightarrow f_{\circ_1}(\lambda) \\ \mathrm{Data}_1 &\rightarrow f_1(\lambda \,|\, \mathrm{Data}_1, f_{\circ_1}(\lambda)) \\ &\rightarrow f_1(r \,|\, \mathrm{Data}_1, \mathcal{L}_1, f_\circ(r))\,. \end{aligned}$$

- The process is iterated for the second experiment:

$$\begin{aligned} f_1(r)\,\&\,\mathcal{L}_2 &\rightarrow f_{\circ_2}(\lambda) \\ \mathrm{Data}_2 &\rightarrow f_2(\lambda \,|\, \mathrm{Data}_2, f_{\circ_2}(\lambda)) \\ &\rightarrow f_2(r \,|\, \mathrm{Data}_2, \mathcal{L}_2, f_1(r)) \\ &\rightarrow f_2(r \,|\, (\mathrm{Data}_1, \mathcal{L}_1), (\mathrm{Data}_2, \mathcal{L}_2), f_\circ(r))\,, \end{aligned}$$

- and so on for all the experiments.

Lets us see in detail the case of null observation in all experiments ($\underline{x} = \underline{0}$) , starting from a uniform distribution.

Experiment 1:

$$\begin{aligned} f_1(\lambda \,|\, x_1 = 0) &= e^{-\lambda} \\ f_1(r \,|\, x_1 = 0) &= \mathcal{L}_1 e^{-\mathcal{L}_1 r} \qquad (9.7) \\ r_{u_1} &= \frac{-\ln 0.05}{\mathcal{L}_1} \text{ at } 95\% \text{ probability}\,. \qquad (9.8) \end{aligned}$$

Experiment 2:

$$\begin{aligned} f_{\circ_2} &= \frac{\mathcal{L}_1}{\mathcal{L}_2} e^{-\frac{\mathcal{L}_1}{\mathcal{L}_2}\lambda} \\ f_2(\lambda \,|\, x_2 = 0) &\propto e^{-\lambda} \frac{\mathcal{L}_1}{\mathcal{L}_2} e^{-\frac{\mathcal{L}_1}{\mathcal{L}_2}\lambda} \\ &\propto e^{-\left(1+\frac{\mathcal{L}_1}{\mathcal{L}_2}\right)\lambda} \\ f_2(r \,|\, x_1 = x_2 = 0) &= (\mathcal{L}_1 + \mathcal{L}_2)\, e^{-(\mathcal{L}_1+\mathcal{L}_2)\, r}. \end{aligned}$$

[2] This two-step inference is not really needed, but it helps to follow the inferential flow. One could think more directly of

$$f(x \,|\, r, \mathcal{L}_i) = \frac{e^{-r\,\mathcal{L}_i}(r\,\mathcal{L}_i)^x}{x!}\,.$$

When the dependence between the two quantities is not linear, a two-step inference may cause trouble: see comments in Section 9.3.3.

Experiment n:

$$f_n(r \,|\, \underline{x} = \underline{0}, f_\circ(r) = k) = \sum_i \mathcal{L}_i e^{-\sum_i \mathcal{L}_i r} \,. \tag{9.9}$$

The final result is insensitive to the data grouping. As the intuition suggests, many experiments give the same result of a single experiment with equivalent luminosity. To get the upper limit, we calculate, as usual, the cumulative distribution and require a certain probability P_u for r to be below r_u [i.e. $P_u = P(r \le r_u)$]:

$$\begin{aligned} F_n(r \,|\, \underline{x} = \underline{0}, f_\circ(r) = k) &= 1 - e^{-\sum_i \mathcal{L}_i r} \\ r_u &= \frac{-\ln(1-P_u)}{\sum_i \mathcal{L}_i} \\ \frac{1}{r_u} &= \frac{-\sum_i \mathcal{L}_i}{\ln(1-P_u)} \\ &= \sum_i \frac{-\mathcal{L}_i}{\ln(1-P_u)} \\ &= \sum_i \frac{1}{r_{u_i}} \,, \end{aligned}$$

obtaining the following rule for the combination of upper limits on rates:

$$\frac{1}{r_u} = \sum_i \frac{1}{r_{u_i}} \,. \tag{9.10}$$

We have considered here only the case in which no background is expected, but it is not difficult to take background into account, following what has been said in Section 5.6.5.

9.1.4 Including systematic effects

A last interesting case is when there are systematic errors of unknown size in the detector performance. Independently of where systematic errors may enter, the final result will be an uncertainty on $\mathcal{L}$. In the most general case, the uncertainty can be described by a probability density function:

$$f(\mathcal{L}) = f(\mathcal{L} \,|\, \text{best knowledge on experiment}) \,.$$

For simplicity we analyse here only the case of a single experiment. In the case of many experiments, we only need to iterate the Bayesian inference, as has often been shown in these notes.

Following the general lines given in Section 2.10.3, the problem can be solved by considering the conditional probability, obtaining :

$$f(r \,|\, \text{Data}) = \int f(r \,|\, \text{Data}, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} \,. \tag{9.11}$$

The case of absolutely precise knowledge of $\mathcal{L}$ is recovered when $f(\mathcal{L})$ is a Dirac delta.

Let us treat in some more detail the case of null observation ($\underline{x} = \underline{0}$). For each possible value of $\mathcal{L}$ one has an exponential of expected value $1/\mathcal{L}$ [see Eq. (9.7)]. Each of the exponentials is weighted with $f(\mathcal{L})$. This means that, if $f(\mathcal{L})$ is rather symmetrical around its barycentre (expected value), in a first approximation the more and less steep exponentials will compensate,

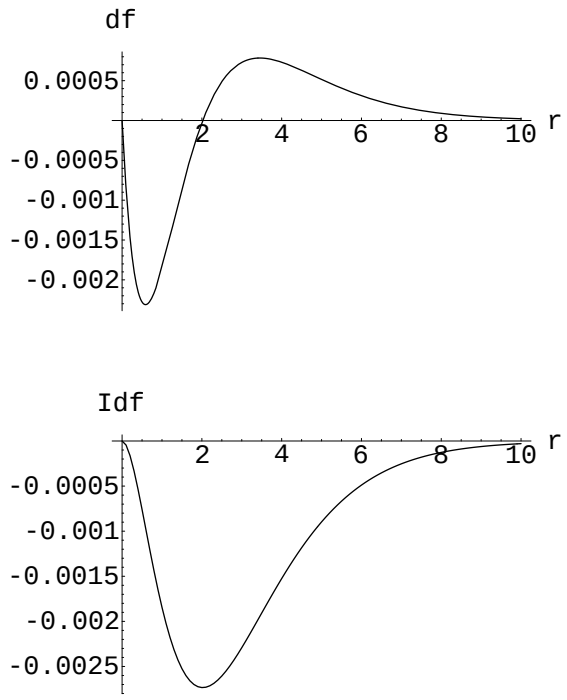

Figure 9.1: Inference on the rate of a process, with and without taking into account systematic effects: upper plot: difference between $f(r \,|\, x = 0, \mathcal{L} = 1.0 \pm 0.1)$ and $f(r \,|\, x = 0, \mathcal{L} = 1 \pm 0)$, using a normal distribution of $\mathcal{L}$; lower plot: integral of the difference, to give a direct idea of the variation of the upper limit.

and the result of integral (9.11) will be close to $f(r)$ calculated in the barycentre of $\mathcal{L}$, i.e. in its nominal value $\mathcal{L}_\circ$:

$$\begin{aligned} f(r \,|\, \text{Data}) &= \int f(r \,|\, \text{Data}, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} \approx f(r \,|\, \text{Data}, \mathcal{L}_\circ) \\ r_u \,|\, \text{Data} &\approx r_u \,|\, \text{Data}, \mathcal{L}_\circ \,. \end{aligned}$$

To make a numerical example, let us consider $\mathcal{L} = 1.0 \pm 0.1$ (arbitrary units), with $f(\mathcal{L})$ following a normal distribution. The upper plot of Fig. 9.1 shows the difference between $f(r \,|\, \text{Data})$ calculated applying Eq. (9.11) and the result obtained with the nominal value $\mathcal{L}_\circ = 1$:

$$\begin{aligned} df &= f(r \,|\, x = 0, f(\mathcal{L})\,) - f(r \,|\, x = 0, \mathcal{L} = 1.0) && (9.12) \\ &= \int f(r \,|\, x = 0, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} - e^{-r} \,. && (9.13) \end{aligned}$$

d is negative up to $r \approx 2$, indicating that systematic errors normally distributed tend to increase the upper limit. But the size of the effect is very tiny, and depends on the probability level chosen for the upper limit. This can be seen better in the lower plot of Fig. 9.1, which shows the integral of the difference of the two functions. The maximum difference is for $r \approx 2$. As far as the upper limits are concerned, we obtain (the large number of — non-significatant—digits

is only to observe the behaviour in detail):

$$\begin{aligned} r_u(x=0, \mathcal{L}=1\pm 0, \text{ at } 90\%) &= 2.304 \\ r_u(x=0, \mathcal{L}=1.0\pm 0.1, \text{ at } 90\%) &= 2.330 \\ r_u(x=0, \mathcal{L}=1\pm 0, \text{ at } 95\%) &= 2.996 \\ r_u(x=0, \mathcal{L}=1.0\pm 0.1, \text{ at } 95\%) &= 3.042\,. \end{aligned}$$

An uncertainty of 10% due to systematics produces only a 1% variation of the limits.

To simplify the calculation (and also to get a feeling of what is going on) we can use some approximations.

1. Since the dependence of the upper limit of r from $1/\mathcal{L}$ is given by

$$r_u = \frac{-\ln(1-P_u)}{\mathcal{L}}\,,$$

the upper limit averaged with the belief on $\mathcal{L}$ is given by

$$r_u = -\ln(1-P_u)\,\mathrm{E}\left[\frac{1}{\mathcal{L}}\right] = \int \frac{1}{\mathcal{L}} f(\mathcal{L})\,\mathrm{d}\mathcal{L}\,.$$

We need to solve an integral simpler than in the previous case. For the above example of $\mathcal{L}=1.0\pm 0.1$ we obtain $r_u = 2.326$ at 90% and $r_u = 3.026$ at 95%.

2. Finally, as a real rough approximation, we can take into account the small asymmetry of r_u around the value obtained at the nominal value of $\mathcal{L}$ averaging the two values of $\mathcal{L}$ at $\pm\sigma_{\mathcal{L}}$ from $\mathcal{L}_\circ$:

$$\begin{aligned} r_u &\approx \frac{-\ln(1-P_u)}{2}\left(\frac{1}{\mathcal{L}_\circ-\sigma_{\mathcal{L}}}+\frac{1}{\mathcal{L}_\circ+\sigma_{\mathcal{L}}}\right) \\ &\approx \frac{-\ln(1-P_u)}{\mathcal{L}_\circ}\left(1+\frac{\sigma^2_{\mathcal{L}}}{\mathcal{L}^2_\circ}\right)\,. \end{aligned}$$

We obtain numerically identical results to the previous approximation.

The main conclusion is that the uncertainty due to systematics plays only a second-order role, and it can be neglected for all practical purposes. A second observation is that this uncertainty increases slightly the limits if $f(\mathcal{L})$ is distributed normally, but the effect could also be negative if the $f(\mathcal{L})$ is asymmetric with positive skewness.

As a more general remark, one should not forget that the upper limit has the meaning of an uncertainty and not of a value of quantity. Therefore, as nobody really cares about an uncertainty of 10 or 20% on the uncertainty, the same is true for upper/lower limits. At the per cent level it is mere numerology (I have calculated it at the 10^{-4} level just for mathematical curiosity).

9.1.5 Is there a signal?

There is an important remark to be made on the interpretation of the result: can we conclude from an upper limit that the searched for signal does not exist? Tacitly yes. But let us take the final distribution of λ for $x=0$ (with a uniform prior and neglecting systematic effects) and let us read the result in a complementary way:

$$P(\lambda \geq \lambda_L) = e^{-\lambda_L}\,.$$

We obtain, for example:

$$\begin{aligned} P(\lambda \geq 10^{-1}) &= 90\% \\ P(\lambda \geq 10^{-2}) &= 99\% \\ \ldots &\quad \ldots \end{aligned}$$

Since $P(\lambda = 0) = 0$, it seems that we are almost sure that there is a signal, although of very small size. The solution to this apparent paradox is to remember that the analysis was done assuming that a new signal existed and that we only wanted to infer its size from the observation, under this assumption. On the other hand, from the experimental result we cannot conclude that the signal does not exist.

For the purpose of these notes, we follow the good sense of physicists who, for reasons of economy and simplicity, tend not to believe in a new signal until there is strong evidence that it exists. However, to state with a number what 'strong evidence' means is rather subjective. For a more extensive discussion about this point see Ref. [25].

9.1.6 Signal and background: a *Mathematica* example

As a last application of the Poissonian model, let us make a numerical example of a counting measurement in the presence of background. To compare full and approximative results, let us choose a number large enough for the normal approximation to be reasonable. For example, we have observed 44 counts with an expected background of 28 counts. What can we tell about the signal? We solve the problem with the following *Mathematica* code[3] applied to the formulae of Section 5.6.5 (s stands for λ_s and b for $\lambda_{B_\circ}$):

```
(*******************************************************)
ClearAll["Global‘*"]

f = (Exp[-s]*(b0+s)^x)/(x!Sum[b0^i/i!, {i, 0, x}])

x=44;
b0=28;

m = NIntegrate[s*f, {s, 0, 1E^6}]
sigma = Sqrt[NIntegrate[s^2*f, {s, 0, 1E^6}] - m^2]

Plot[f, {s, 0, 50}, AxesLabel->{s, "f"}]

fd1=D[Log[f], s];
fd2= D[fd1, s];

res=FindMinimum[-f, {s,m}];
smax = res[[2]]
sigma2=1/Sqrt[-(fd2 /. res[[2]])]
(*******************************************************)
```

The code evaluates and plots the final distribution of λ_s obtained from a uniform prior [formula (5.88)] and calculates:

- the prevision $\mathrm{E}(\lambda_S)$

$$\mathtt{m} \equiv \mathrm{E}(\lambda_S) = 17.0\,;$$

[3] If you are interested in Bayesian analysis with *Mathematica* you may take a look at Refs. [81] and [82] (I take for responsibility on the quality of the products, as I have never used them).

- the standard deviation $\sigma(\lambda_S)$:

$$\texttt{sigma} \equiv \sigma(\lambda_S) = 6.7\,;$$

- the mode λ_{S_m}:

$$\texttt{smax} \equiv \lambda_{S_m} = 16.0\,;$$

- the approximated standard deviation calculated from the shape of the final distribution around the mode (see Section 2.9):

$$\texttt{sigma2} = 6.6\,.$$

The resulting probability density function for the signal is shown in Fig. 9.2.

The approximate, but still Bayesian, reasoning to get the same result is as follows.

1. Given this status of information, the certain quantities are:
   - The average value of the background: $\lambda_{B_\circ} = 28\pm \approx 0$;
   - The observation $x = 44$ (it does not even make sense to write ± 0: 44 is 44 !).
2. Instead, we are uncertain on the parameter λ of the Poissonian distribution responsible for the observed number of counts; we can infer [see (5.59) and (5.61)]

$$\lambda \approx x \pm \sqrt{x} = 44.0 \pm 6.6\,.$$

3. Since λ is due to the contribution of the signal and background, we have, finally:

$$\lambda_S = \lambda - \lambda_{B_\circ} = 16.0 \pm 6.6\,.$$

The last evaluation is an example of how Bayesian reasoning helps, independently of explicit use of Bayes' theorem. Nevertheless, these results are still conditioned by the assumption that the signal looked for exists. In fact, Fig. 9.2 does not really prove, from a logical point of view, that the signal does exist, although the distribution seems so nicely separated from zero (see also Ref. [25]).

9.2 Unbiased results

In the Bayesian approach there is a natural way of giving results in an unbiased way, so that everyone may draw his own scientific conclusion depending on his prior beliefs. One can simply present likelihoods or, for convenience, ratios of likelihoods (Bayes' factors, see Sections 3.5 and 8.8). Some remarks are needed in order not to give the impression that, at the end of this long story, we have not just ended up at likelihood methods.

- First, without priors, the likelihoods cannot be turned into probabilities of the values of physics quantities or of probabilities of hypotheses. Even the 'mathematically harmless' uniform distribution, which gets simplified in Bayes' formula, does its important job. For this reason publishing only likelihoods does not mean publishing unbiased conclusions, but rather publishing no conclusions! Hence, one is not allowed to use this 'result' for uncertainty propagation, as it has no uncertainty meaning.

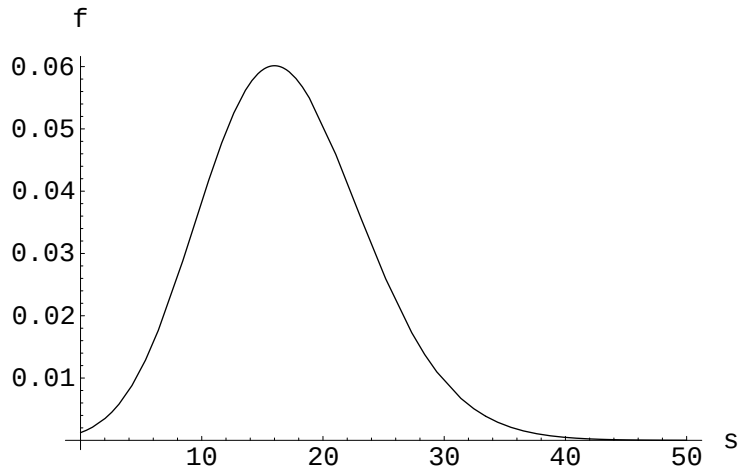

Figure 9.2: Final distribution for the signal ($s \equiv \lambda_S$), having observed 44 counts, with an expected background of 28 counts.

- This game of avoiding the priors can be done only at the final level of the analysis. Presuming priors can be avoided for each step is absurd, in the sense that the reader has no way of completely redoing the analysis by plugging in his preferred priors at all the crucial points. If the experimenter refrains to choose a prior, but, nevertheless, he goes on in the steps of the analysis, he is in fact using uniform priors in all inferences. This may be absolutely reasonable but one has to be aware of what one is doing. If instead there are reasons for using non-uniform priors in some parts of the analysis, as his past experience suggested, the experimenter should not feel guilty. There are so many subjective and even really arbitrary ingredients in a complex analysis that we must admit, if we believe somebody's results and we use his conclusions as if they were our conclusions, it is simply because we trust him. So we are confident that his knowledge of the detector and of the measurement is superior to ours and this justifies his choices. As a matter of fact, the choice of priors is insignificant compared with all the possible choices of a complicated experiment.

- The likelihoods are probabilities of observables given a particular hypothesis. Also their evaluation has subjective (and arbitrary) contributions. Sticking to the idealistic position of providing only objective data is equivalent to stagnating research.

Having clarified these points, let us look at two typical cases.

Classifying hypotheses. In the case of a discrete number of hypotheses, the proper quantities to report are the likelihoods of the data for each hypothesis

$$P(\text{data}\,|\,H_i)\,,$$

or Bayes' factor for any of the couples

$$\frac{P(\text{data}\,|\,H_i)}{P(\text{data}\,|\,H_j)}\,.$$

On the other hand, the likelihood for a given hypothesis alone, e.g. $P(\text{data}\,|\,H_\circ)$, does not help the reader to form his idea on the hypothesis, nor on alternatives (see also Section

8.8). Therefore, if a collaboration publishes experimental evidence against the Standard Model, suggesting some kind of explanation in terms of a new effect, it should report the likelihoods for both hypotheses. (See also 5th bullet of Section 8.8 in the case of Gaussian likelihood).

Values of quantities. In this case the likelihoods are summarized by the likelihood function $f(\text{data}\,|\,\mu)$. In this case one may also calculate Bayes' factors between any pair of values

$$\frac{f(\text{data}\,|\,\mu_i)}{f(\text{data}\,|\,\mu_j)}\,.$$

This can be interesting if only a discrete number of solutions are admissible.

When one publishes a likelihood function this should be clearly stated. Otherwise the temptation to turn $f(\text{data}\,|\,\mu)$ into $f(\mu\,|\,\text{data})$ is really strong. In fact, taking the example of the neutrino mass of Section 1.7, the formula

$$\frac{1}{\sqrt{2\,\pi}\,2}\,e^{-\frac{(-4-\mu)^2}{8}}$$

(with mass in eV) can easily be considered as if it were a result for μ:

$$\frac{1}{\sqrt{2\,\pi}\,2}\,e^{-\frac{[\mu-(-4)]^2}{8}}$$

and conclude that $\mu_\nu = -4 \pm 2\,\text{eV}$.

After having criticized this way of publishing the data for the second time, I try in the next section to encourage this way of presenting the result, on condition that one is well aware of what one is writing.

9.2.1 Uniform prior and fictitious quantities

Let us consider n independent data sets, or experiments, each of which gives information on the quantity μ. For each data set there is a likelihood

$$f_i(\text{data}_i\,|\,\mu)\,.$$

Each data set gives, by itself, the following information:

$$f(\mu\,|\,\text{data}_i) \quad \propto \quad f_i(\text{data}_i\,|\,\mu)\cdot f_\circ(\mu)\,. \tag{9.14}$$

The global inference is obtained using Bayes' theorem iteratively:

$$f(\mu\,|\,\bigcup_i \text{data}_i) \quad \propto \quad \Pi_i f_i(\text{data}_i\,|\,\mu)\cdot f_\circ(\mu)\,. \tag{9.15}$$

We may use, as a formal tool, a ficticious inference $\tilde{f}(\mu)$ using for each data set a uniform prior in the range $-\infty < \mu + \infty$:

$$\tilde{f}_i(\mu\,|\,\text{data}_i) \quad \propto \quad f_i(\text{data}_i\,|\,\mu)\cdot k\,. \tag{9.16}$$

This allows us to rewrite

$$f(\mu\,|\,\bigcup_i \text{data}_i) \quad \propto \quad \Pi_i \tilde{f}_i(\mu\,|\,\text{data}_i)\cdot f_\circ(\mu)\,.$$

This stratagem has the advantage that one can report 'pseudoresults' on fictitious quantities which, in the case of Gaussian likelihoods, may be combined according to the usual formula of the average with the inverse of the variances (see Section 5.4.2). They can be transformed, finally, into the physical result using the physical prior $f_\circ(\mu)$. It is important to state the procedure clearly and, if possible, to indicate the fictitious quantity with different symbols. For example, the result of the problem of Section 5.4.3 can be reported in the following way:

> "From the observed value of –5.4 eV and the knowledge of the likelihood, described by a normal distribution centred in the true value of the mass with $\sigma = 3.3\,\mathrm{eV}$ independent of the mass, we get a fictitious mass of
>
> $$\tilde{m}_\nu = -5.4 \pm 3.3\,\mathrm{eV}\,,$$
>
> where 'fictitious' indicates a hypothetical mass which could assume any real number with uniform distribution. Assuming the more physical hypothesis $f_\circ(m_\nu) \geq 0$ yields to ... (see figure ...), from which follows a 95% upper limit of 3.9 eV."

The conclusion of this section is that the uniform prior is a convenient prior for many purposes:

- it produces results very similar to those obtainable using the rational priors of those who have done the experiment, as shown in many of the examples given in these notes (see, for example, Section 5.4.3);
- it allows easy combination of data and a physics motivated prior can be added at the end;
- there is no problem of 'double counting' the same prior, as would happen if several experimenters were to use the same non-uniform prior to infer the same quantity from different data.

The problem of presenting unbiased results in frontier measurements is also discussed in Refs. [26], [25], [83] and [84].

9.3 Constraining the mass of a hypothetical new particle: analysis strategy on a toy model

As a last example of an application, let us consider a case which somehow reminds one of the current effort to reduce the uncertainty of the mass of the Higgs particle. Since I don't have access to the original data, and I don't want this exercise to be considered as any a kind of claim, I will just invent the rules of the game.[4] So, physics data and results will be imaginary, but the inferential procedure will be performed according to what I consider to be the proper way of doing it.

9.3.1 The rules of the game

The hypothetical world of this analysis is:

[4] This section is intentionally pedagogical. An analysis using the best physical assumptions can be found in Ref. [26]. Indeed, this analysis follows the strategy outlined here, with some variations introduced to match the information available in the real situation.

- three experiments (A, B and C) took data in an e^+e^- collider, at different energies and with different sensitivity to the H particle production ('H stands for hypothetical...'). The experiments reported 0 candidates and no background was expected (this is a minor approximation to simplify the formulae: we have seen how the background and its uncertainty may be treated).
- The beam energy was 0.09 and 0.1 in arbitrary units (you may think of TeV) and the kinematical factor which suppresses the production near threshold (and eventually takes into account efficiencies, tagging, etc.) is chosen somehow 'arbitrarily' to be β^3 factor, where β is the velocity of the pair produced particles.
- Cross-section and integrated luminosity are summarized into a sensitivity factor k, such that the expected number of events is

$$\lambda = k\,\beta^3 = k\,\left(1 - \frac{m^2}{E_b^2}\right)^{3/2} .$$

- We also have other pieces of information on H: two indirect determinations are characterized by a Gaussian likelihood, and each of them would allow a Gaussian determination of the mass, if one considered that this could be uniformly distributed from $-\infty$ to $+\infty$ (see Section 9.2).
- The five datasets are considered to be independent.
- The prior of the scientific community about the value of the mass has changed in recent years, due not only to negative results, but also to theoretical progress:
  - essentially, once there was uncertainty even in the order of magnitude, i.e. $f(\ln m) \approx k$, yielding $f(m) \approx 1/m$; as a conservative position, one could still stick to this position;
  - at present, many think that $\mathcal{O}(m) \approx 0.1$–$0.2$; this state of uncertainty can be modelled by a uniform distribution over the range of interest.

9.3.2 Analysis of experiment A

A has been run at a c.m. energy of 2×0.09, with sensitivity factor 20. It has observed 0 H candidate events. One can proceed in two ways, one correct and the other wrong. Let us start with the wrong one.

9.3.3 Naïve procedure

We have learned that we can infer

$$f(\lambda \,|\, x = 0) = e^{-\lambda}$$

and, from the relation $\lambda = k\,(1 - m^2/E_b^2)^{3/2}$ we get

$$f(m) = 3\,k\,\frac{m}{E_b^2}\left(1 - \frac{m^2}{E_b^2}\right)^{\frac{1}{2}} \exp\left[-k\,\left(1 - \frac{m^2}{E_b^2}\right)^{\frac{3}{2}}\right], \tag{9.17}$$

which is clearly wrong, since it goes to 0 for $m \to E_b$. Figure 9.3 shows the plot of $f(m)$, obtained with the following *Mathematica* code:

```
(*******************************************************)
ClearAll["Global`*"]

(* Experiment A has been run at beam energy 0.09, with
sensitivity factor k=20, threshold function beta^3,
and has observed 0 events.*)

ka=20
eba=0.09

v=Sqrt[1-(m/eba)^2]
lambda= ka*v^3

(* f(m) obtained from f(lambda)=Exp[-lambda] by lambda=k beta^3
   (threshold factor) using p.d.f transformation *)

fl=Exp[-lambda]
J=Abs[D[lambda, m]]
fm=fl*J

(* check normalization and plot *)

NIntegrate[fm, {m, 0, eba}]
Plot[fm, {m, 0.06, eba}, AxesLabel -> {m, f}]

(* Strange result: try to figure out the reason! *)
(*******************************************************)
```

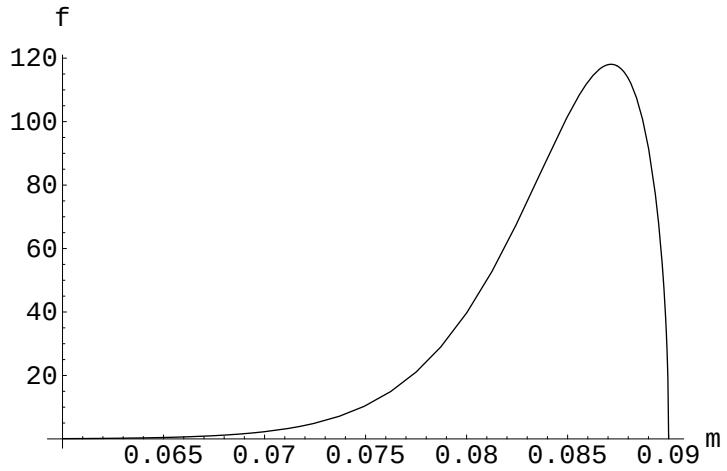

Figure 9.3: Inference on the mass of the hypothetical particle H with the information of experiment A, obtained from the intermediate inference on λ assuming a uniform prior on this quantity. The result looks very strange.

The result is against our initial rational belief and we refuse to accept it. The origin of this strange behaviour is due to the term $\sqrt{1-m^2/E_b^2}$ in (9.17) that comes directly from the Jacobian of the transformation and, indirectly from having assumed a prior uniform in λ. To solve the problem we have to change prior. But this seems to be cheating. Should not the prior come before? How can we change it after we have seen the final distribution?

We should not confuse what we think with how we model it. The intuitive prior is on the mass values, and this prior should be flat (at least at this stage of the analysis, as discussed in Section 9.2). Unfortunately, what is flat in λ is not flat in m, and vice versa. This problem has been discussed in Section 5.3. In fact, it is not really a problem of probability, but of extrapolating intuitive probability (which is at the basis of subjective probability and only deals with discrete numbers) to continuous variables. This is the price we pay for using all the mathematical tools of differential calculus. But one has to be very careful in formulating the problem. If one wants to get rid of these problems, one may discretize λ and m in a way which is consistent with to the experimental resolution. If we discretize, a flat distribution in m is mapped to a flat distribution in λ. And the problems caused by the Jacobian go away with the Jacobian itself, at the expense of some complications in computation.

9.3.4 Correct procedure

In order to solve the problem consistently with our beliefs, we have to avoid the intermediate inference[5] on λ, and write prior and likelihood directly in terms of m:

$$f(m \,|\, x=0) \propto \exp\left[-k\,\left(1-\frac{m^2}{E_b^2}\right)^{\frac{3}{2}}\right] \cdot f_\circ(m)\,, \tag{9.18}$$

with $f_\circ(m) = \textit{constant}$. Let us do it again with *Mathematica*:

```
(*******************************************************)
(* Now let's do it right: *)

lik=Exp[-lambda]
norm=NIntegrate[lik, {m, 0, eba}]

(* fa(m) is the final distribution from experiment A,
   under the condition that m < eba *)

fa=lik/norm
Plot[fa, {m, 0.06, eba}, AxesLabel -> {m, f}]
(*******************************************************)
```

The final distribution is shown in Fig. 9.4. It is now reasonable and consistent with the expectations: The values of mass which are less believable are those which could have been produced easier, given the kinematics. From $f(m \,|\, x=0)$ we can calculate several results, for example a 95% upper limit, the average and the standard deviation:

```
(*******************************************************)
NIntegrate[fa, {m, 0, 0.0782}]
ava = NIntegrate[m*fa, {m, 0, eba}]
stda = Sqrt[NIntegrate[m*fa, {m, 0, eba}] - ava^2]
(*******************************************************)
```

We get:

$$m \;>\; 0.0782 \text{ with } 95\% \text{ probability} \tag{9.19}$$

$$\mathrm{E}(m) \;=\; 0.0856 \tag{9.20}$$

$$\sigma(m) \;=\; 0.0038\,. \tag{9.21}$$

[5] A two-step inference was shown in Section 9.1.3 for the case of monopole search. There there was no problem because λ and r are linearly related.

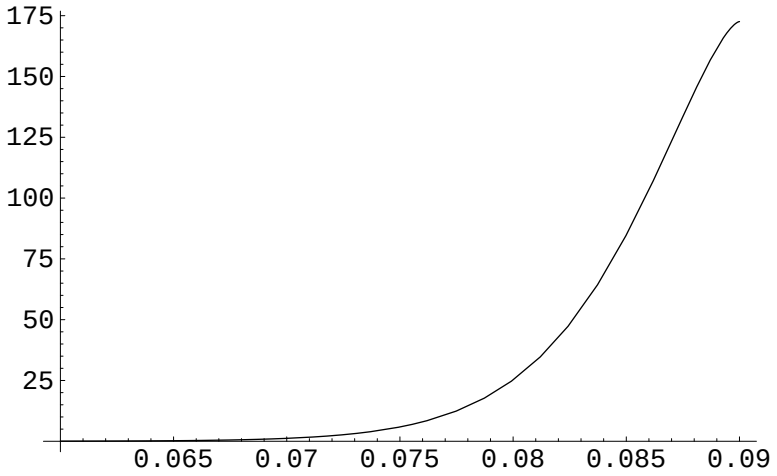

Figure 9.4: Inference on m obtained from a direct inference on m, starting from a uniform prior in this quantity.

9.3.5 Interpretation of the results

At this point we need to discuss the results achieved so far, to understand what they mean, and what they don't.

- Can we say that we believe that m is above 0.0782 with 95% probability and below with 5%? Would you bet 5 \$ that m is below 0.0782 with the hope of winning 100 \$ if it turns out to be the case?
- Can we say we believe that we have done a precise measurement of m, since it looks like $m = 0.0856 \pm 0.0038$?
- How can we say that $m = 0.0856 \pm 0.0038$ and speak about a lower bound?

In particular, the first statement gives the impression that we can say something about the mass even if H was too heavy to be produced. In general, a statement like

> *"A 95% confidence level lower bound of* $77.5\,GeV/c^2$ *is obtained for the mass of the Standard Model Higgs boson."* [85]

may be misleading, because it transmits information which is inconsistent with the experimental observation. The interpretation of the result (9.19) is limited to

$$P(m > 0.0782 \,|\, \underline{0 \leq m \leq 0.09}) \quad = \quad 0.95\,, \tag{9.22}$$

as may be understood intuitively and will be shown in a while (there is also the condition of the uniform prior, but at this level it is irrelevant). So, given this status of information, I could bet 1:19 that the m is below 0.0782, but only on condition that the bet is invalidated if m turns out to be greater than the beam energy (see Section 3.4.2). Otherwise, I would choose the other direction (19:1 on '$m > 0.0782$') without hesitation (and wish fervently that somebody accepts my bet ...).

What are our rational beliefs on m, on the basis of experiment A, releasing the condition $\leq m \leq E_b$? The data cannot help much because there is no experimental sensitivity, and the conclusions depend essentially on the priors.

To summarize, the result of the inference is:

$\mathbf{m < E_b}$**:** $P(m > 0.0782) = 0.95$; $m = 0.0856 \pm 0.038$, etc. ;

$\mathbf{m \geq E_b}$**:** "HIC SUNT LEONES"[6]

As a final remark on the presentation of the result, I would like to comment on the three significative digits with which the result on the 'conditional lower bound' has been given. For the sake of the exercise the mass bound has been evaluated from the condition (9.22). But does it really matter if the limit is 0.0782, rather than 0.0780, or 0.0800? As stated in Sections 5.4.3 and 9.1.1, the limits have to be considered in the same way as the uncertainty. Nobody cares if the uncertainty of the uncertainty is 10 or 20%, and nobody would redo a MACRO-like experiment to lower the monopole limit by 20%. Simply translating this argument to the case under study, it may give the impression that one significant digit would be enough (0.08), but this is not true, if we stick to presenting the result under the condition that m is smaller than E_b. In fact, what really matters, is not the absolute mass, but the mass difference with respect to the kinematical limit. If the experiment ran with infinite statistics and found 'nothing', there is no interest in providing a detailed study for the limit: it will be exactly the same as the kinematical limit. Therefore, the interesting piece of information that the experimenter should provide is how far the lower bound is from the kinematical limit, i.e. what really matters is not the absolute mass scale, but rather the mass difference. In our case we have

$$\Delta m = E_b - \textit{lower bound} = 0.0118 \rightarrow 0.012\,. \tag{9.23}$$

Two digits for this number are enough (or even only one, if the first were greater that 5) and the value of the lower bound becomes[7]

$$m > 0.078 \text{ at } 95\%, \text{ if } 0 \leq m \leq 0.09\,.$$

9.3.6 Outside the sensitivity region

With the Bayesian method it is possible to trace the point in which an unstated condition has been introduced, and how to remove it, or how to take it into account. With the form of the likelihood used in (9.18) it was implicit that m should not exceed E_b. A more physically motivated likelihood should be:

$$f(x = 0 \,|\, m) = \begin{cases} \exp\left[-k\,\left(1 - \frac{m^2}{E_b^2}\right)^{3/2}\right] & \text{if} \quad 0 \leq m \leq E_b \\ 1 & \text{if} \quad m > E_b \end{cases} \tag{9.24}$$

Taking a uniform prior, we get the following posterior:

$$f(m \,|\, x = 0) = \begin{cases} \dfrac{\exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right]}{\int_0^{E_b} \exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right] \mathrm{d}m + (m_{max} - E_b)} & \text{if} \quad 0 \leq m \leq E_b \\ \dfrac{m_{max} - E_b}{\int_0^{E_b} \exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right] \mathrm{d}m + (m_{max} - E_b)} & \text{if} \quad m > E_b \end{cases} \tag{9.25}$$

[6] *"Here are the lions"* is what the ancient Romans used to write on the parts of their maps representing unexplored regions.

[7] Numerologists may complain that this does not correspond to exactly 95%, but the same happens when a standard uncertainty is rounded to one or two digits and the probability level calculated from the rounded number may differ a lot from the nominal 68.3% calculated from the original value. But who cares?

where $(m_{max} - E_b)$ comes from the integral $\int_{E_b}^{m_{max}} 1 \cdot \mathrm{d}m$. So, we get our solution (9.18) for $m_{max} = E_b$. In general, the probability that $m \leq E_b$ is smaller than 1 and decreases for increasing m_{max}. For the parameters of experiment A the integral in the denominator is equal to 0.0058. Therefore, if, for example, $m_{max} = 3\,E_b$

$$\begin{aligned} P(m < E_b \,|\, x = 0, m_{max} = 3\,E_b) &= 2.7\% \\ P(m < 0.078 \,|\, x = 0, m_{max} = 3\,E_b) &= 0.13\%. \end{aligned}$$

There is another reasoning which leads to the same conclusion. At $m = E_b$ the detector has zero sensitivity. For this reason, in case of null observation, this values gets the maximum degree of belief. As far as larger values are concerned, the odds ratios with respect to $m = E_b$ must be invariant, since they are not influenced by the experimental observations, i.e.

$$\frac{f(m \,|\, x = 0)}{f(m = E_b \,|\, x = 0)} = \frac{f_\circ(m)}{f_\circ(m = E_b)} \qquad (m > E_b)\,. \tag{9.26}$$

Since we are using, for the moment, a uniform distribution, the condition gives:

$$f(m \,|\, x = 0) = f(m = E_b \,|\, x = 0) \quad (m > E_b)\,. \tag{9.27}$$

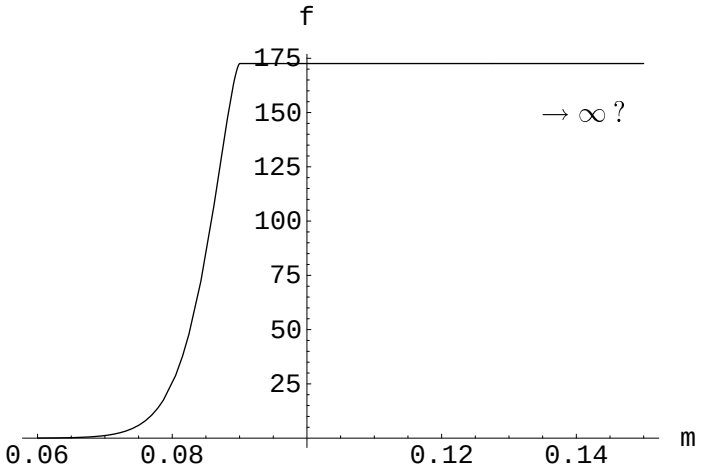

Figure 9.5: Result of the inference from experiment A, taking into account values of mass above the beam energy as well. These all have the same degree of belief and the normalization constant depends on the maximum value of m considered. Therefore the distribution is usually improper.

We easily get the result shown in Fig. 9.5 by this piece of *Mathematica* code:

```
(*******************************************************)
famax=fa/.m->eba
f2a=If[m<eba, fa, famax]

(* f2a(m) represents instead the (improper) distribution
   extended also for values larger that eba, in the light
   of a flat prior and of the Experiment A *)

Plot[f2a, {m,0.06,0.15}]
(*******************************************************)
```

The curve is extended on the right side up to a limit which cannot be determined by this experiment, it could virtually go to infinity. For this reason the ratio of probabilities

$$\frac{P(m < E_b)}{P(m \geq E_b)}$$

decreases (i.e. we tend to believe more strongly large mass values) but its exact value is not well defined. For this reason we leave the function 'open' on the right side and unnormalized. The normalization will be done when we can include other data which can provide an upper limit.

9.3.7 Including other experiments

Each of the other experiments are treated in exactly the same way. Comparing B and C it is interesting to see how the beam energy and the sensitivity factor contribute to constraining the mass. For reasons of space the plots are not shown. This is the rest of the *Mathematica* code to conclude the analysis:

```
(*******************************************************)
(* Experiment B has been run at beam energy 0.09, with
   sensitivity factor k=100,  threshold function beta^3, and
   has observed 0 events.*)

kb=100
ebb=0.09
v=Sqrt[1-(m/ebb)^2]
lambda= kb*v^3
lik=Exp[-lambda]
norm=NIntegrate[lik, {m, 0, ebb}]
fb=lik/norm
avb = NIntegrate[m*fb, {m, 0, ebb}]
Plot[fb, {m, 0.07, ebb}, PlotRange->{0, 600},
     AxesLabel -> {m, f}]

fbmax=fb/.m->ebb
f2b=If[m<ebb, fb, fbmax]
Plot[f2b, {m,0.07,0.15},  PlotRange->{0, 600},
     AxesLabel -> {m, f}]

(* The conclusions from A + B are, with and without the condition m<ebeam,
   respectively (remember that the latter is improper): *)

fcom1ab=fa*fb/NIntegrate[fa*fb, {m, 0, eba}]
avab = NIntegrate[m*fcom1ab, {m, 0, eba}]
Plot[fcom1ab, {m, 0.07, eba},  PlotRange->{0, 600},
     AxesLabel -> {m, f}]
fcom2ab=f2a*f2b

(* Experiment C has been run at beam energy 0.1, with sensitivity factor k=10,
   threshold function beta^3 and, has observed 0 events. *)

kc=10
ebc=0.1
v=Sqrt[1-(m/ebc)^2]
lambda= kc*v^3
lik=Exp[-lambda]
```

```
norm=NIntegrate[lik, {m, 0, ebc}]
fc=lik/norm
Plot[fc, {m, 0.07, ebc}, PlotRange->{0, 100},
     AxesLabel -> {m, f}]
avc = NIntegrate[m*fc, {m, 0, ebc}]
fcmax=fc/.m->(ebc-0.000001)
f2c=If[m<ebc, fc, fcmax]
Plot[f2c, {m,0.07,0.15},  PlotRange->{0, 100},
     AxesLabel -> {m, f}]
(*******************************************************)
```

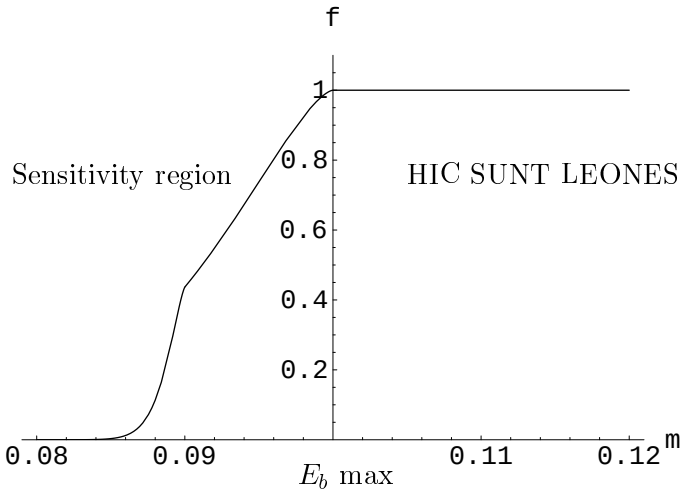

Figure 9.6: Final distribution (improper, see text) on m from experiments A, B and C. The curve has been arbitrarily rescaled to have the maximum of 1.

The combination of the result is done in the usual way, multiplying the likelihoods or the final p.d.f.'s, if these were obtained from a uniform distribution. We only see the combination of the three experiments, shown in Fig. 9.6. Finally, the indirect determinations are also included.

```
(*******************************************************)
(* Conclusions from A + B + C , with and without the condition m<ebeam,
   respectively (remember that the latter is improper): *)

fcom1abc=f2a*f2b*fc/NIntegrate[f2a*f2b*fc, {m, 0, ebc}]
avabc=NIntegrate[m*fcom1abc, {m, 0, ebc}]
Plot[fcom1abc, {m, 0.07, ebc},  PlotRange->{0, 150},
     AxesLabel -> {m, f}]
fcom2abc=f2a*f2b*f2c

(* Now we add independent determinations of m,
   deriving from normal likelihoods,
   and assuming uniform prior *)

g1=1/sigma1/(Sqrt[2*Pi])*Exp[-(m-mu1)^2/2/sigma1^2]
g2=1/sigma2/(Sqrt[2*Pi])*Exp[-(m-mu2)^2/2/sigma2^2]

mu1=0.09
sigma1=0.04
```

```
mu2=0.15
sigma2=0.04

(* The two overall (improper) priors may be a uniform,
   or 1/m, i.e.  flat in ln(m), to express initial
   uncertainty on the order of magnitude of m *)

p1=1
p2=1/m

final1=fcom2abc*g1*g2*p1/NIntegrate[fcom2abc*g1*g2*p1, {m, 0, 10}]
mean1=NIntegrate[m*final1, {m, 0, 10}]
std1=Sqrt[NIntegrate[m^2*final1, {m, 0, 10}]-mean1^2]
Plot[final1, {m, 0.0, 0.25},  PlotRange->{0, 20},
     AxesLabel -> {m, f}]

final2=fcom2abc*g1*g2*p2/NIntegrate[fcom2abc*g1*g2*p2, {m, 0, 10}]
mean2=NIntegrate[m*final2, {m, 0, 10}]
std2=Sqrt[NIntegrate[m^2*final2, {m, 0, 10}]-mean2^2]
Plot[final2, {m, 0.0, 0.25},  PlotRange->{0, 20},
     AxesLabel -> {m, f}]
(*******************************************************)
```

Finally, the two extra pieces of information enable us to constrain the mass also on the upper side and to arrive at a proper distribution (see Fig. 9.7), under the condition that H exists.

From the final distribution we can evaluate, as usual, all the quantities that we find interesting to summarize the result with a couple of numbers. For a more realistic analysis of this problem see Ref. [26].

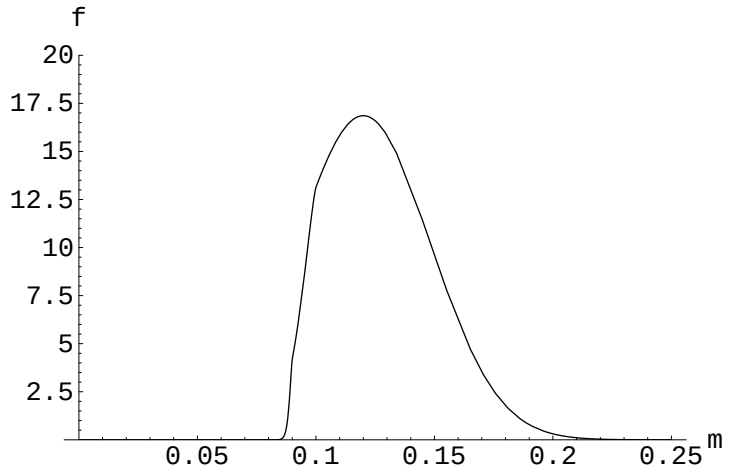

Figure 9.7: Final mass distribution using all five pieces of experimental information, and assuming uniform priors. The curve obtained from the prior $1/m$ does not differ substantially from this.